

Sequences of regressions and their independences

NANNY WERMUTH

Department of Mathematics, Chalmers Technical University, University of Gothenburg

KAYVAN SADEGHI

Department of Statistics, University of Oxford

ABSTRACT: Ordered sequences of univariate or multivariate regressions provide statistical models for analysing data from randomized, possibly sequential interventions, from cohort or multi-wave panel studies, but also from cross-sectional or retrospective studies. Conditional independences are captured by what we name regression graphs, provided the generated distribution shares some properties with a joint Gaussian distribution. Regression graphs extend purely directed, acyclic graphs by two types of undirected graph, one type for components of joint responses and the other for components of the context vector variable. We review the special features and the history of regression graphs, derive criteria to read all implied independences of a regression graph and prove criteria for Markov equivalence that is to judge whether two different graphs imply the same set of independence statements. Knowledge of Markov equivalence provides alternative interpretations of a given sequence of regressions, is essential for machine learning strategies and permits to use the simple graphical criteria of regression graphs on graphs for which the corresponding criteria are in general more complex. Under the known conditions that a Markov equivalent directed acyclic graph exists for any given regression graph, we give a polynomial time algorithm to find one such graph.

Key words: Chain graphs, Concentration graphs, Covariance graphs, Graphical Markov models, Independence graphs, Intervention models, Labeled trees, Lattice conditional independence models, Structural equation models.

1 Introduction

A common framework to model, analyse and interpret data for several, partially ordered joint or single responses is a sequence of multivariate or univariate regressions. Each response is to be generated by a set of its regressors, called its *directly explanatory variables*. Based on statistical analyses, one is to decide which of the variables in the

past are needed for this. Thus, for each response, a first ordering determines what is in its past, hence potentially explanatory, and what is in its future, hence never directly explanatory. Furthermore, no variable is taken to be explanatory for itself.

Corresponding *regression graphs* consist of *nodes* and of *edges coupling distinct nodes*. The *nodes represent the variables* and the *edges stand for conditional associations*, directed or undirected. The directly explanatory variables for an individual response variable Y_i show in the graph as the set of nodes from which arrows start and point to node i . These nodes are commonly named the *parents of node i* .

Every missing edge corresponds to a conditional independence constraint. Edges are *arrows for directed associations* and *lines for undirected associations* among *variables on equal standing*, that is among components of joint responses or context variables. A given regression graph often reflects a particular type of study which may be a simple experiment, a more complex intervention or an observational study.

One of the common features of pure experiments and of sequences of interventions with randomized, proportional allocation of individuals to treatments, is that, by study design, some variables can be regarded to act just like independent random variables. For instance, in an experiment with proportional numbers of individuals assigned randomly to each level combination of several experimental conditions, the set of explanatory variables contains no edge in the corresponding regression graph, reflecting a situation of mutual independence. Similarly, with fully randomized interventions, each treatment variable has exclusively arrows starting from its node but no incoming arrow. After statistical analysis, some conditional independences may be appropriate additional simplifications which show as further missing edges.

Sequences of interventions give a time ordering for some of the variables. A time order is also present in cohort or multi-wave panel studies and in retrospective studies which focus on investigating effects of variables at one fixed time point in the past, without the chance of intervening. By contrast, in a strictly cross-sectional study, in which observations for all variables are obtained at the same time, any particular variable ordering is only assumed rather than implied by actual time.

The node set is at the planning stage of empirical studies ordered into ordered sequences of single or joint responses, $Y_a, Y_b, Y_c \dots$ that we call *blocks of variables on equal standing* and draw them in figures as *boxes*. This determines for the following statistical analyses that within each block, there are undirected edges and between blocks there are directed edges, the arrows. The first block on the left contains the *primary responses* of Y_a and the last block on the right contains *context variables, also named the background variables*. After statistical analyses, arrows may start from nodes within any block but always end at a node in one of the blocks in the future.

Thus, there are no arrows pointing to context variables and all arrows point in the same direction, from left to right. An *intermediate variable* is a response to some variables and also explanatory for other variables so that it has both incoming and outgoing arrows in the regression graph.

As an example, we take data from a retrospective study with 283 adult females answering questions about their childhood; see Hardt et al. (2008) when visiting their general practitioner, mostly for some minor health problems. A well-fitting graph is shown in Figure 1. It contains two binary variables, A, B and six quantitative variables. Except for the directly recorded feature age in years, all other variables are derived from answers to questionnaires, coded so that high values correspond to high scores.

The three blocks a, b, c reflect here a time-ordering of vector variables, Y_a, Y_b, Y_c with Y_a representing the joint response of primary interest, Y_b an intermediate vector variable and Y_c a context vector variable. The three individual components of the primary response Y_a are different aspects of how the respondent recollects aspects of her relationship to the mother. The intermediate variable Y_b has two components which reflect severe distress during childhood. The three components of the context variable Y_c capture background information about the respondent and about her family.

The graph of Figure 1, derived after statistical analyses, shows among other independences that Y_a is conditionally independent of Y_c given Y_b , written compactly in terms of sets of nodes as $a \perp\!\!\!\perp c | b$. None of the components of Y_c has an arrow pointing directly to a component of Y_a , but sequences of arrows lead indirectly from c to a via b .

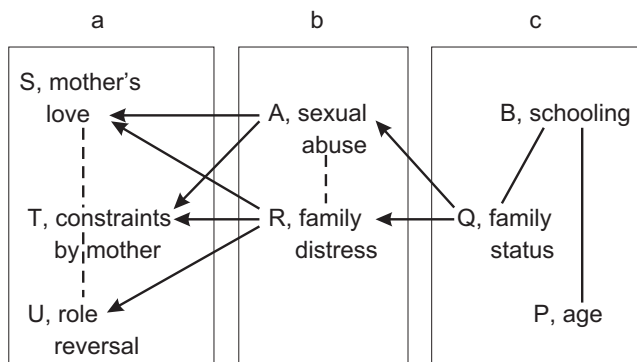


Figure 1: A well-fitting regression graph for data on $n = 283$ adult females; within boxes are Y_a, Y_b, Y_c ; corresponding ordered partitioning of the node set on top of the boxes.

This just says that prediction of Y_a is not improved by knowing the context variable Y_c if information on the more recent intermediate variable Y_b is available. More interpretations of the independences are given later. When the edges present represent substantial associations, the graph may also be viewed as a research hypothesis, the goodness-of-fit

of which can be tested in future studies; see Wermuth and Lauritzen (1990).

Two models are *Markov equivalent* whenever their associated graphs capture the same *independence structure*, that is the graphs lead to the same set of implied independence statements. Markov equivalent models cannot be distinguished on the basis of statistical goodness-of-fit tests for any given set of data. This may pose a problem in machine learning contexts. More precisely, knowledge about Markov equivalent models is essential for designing search procedures which converge with an increasing sample size to a true generating graph; see Castelo and Kocka (2003) for searches within the class of *directed acyclic graphs* which consist exclusively of arrows and capture independences of ordered sequences in single response regressions.

More importantly though, Markov equivalent models may offer alternative interpretations of a given well-fitting model or open the possibility of using different types of fitting algorithms. As we shall see in Section 7, the graph for nodes A, R, B, P, Q in blocks b and c of Figure 1 is Markov equivalent to both graphs of Figure 2.

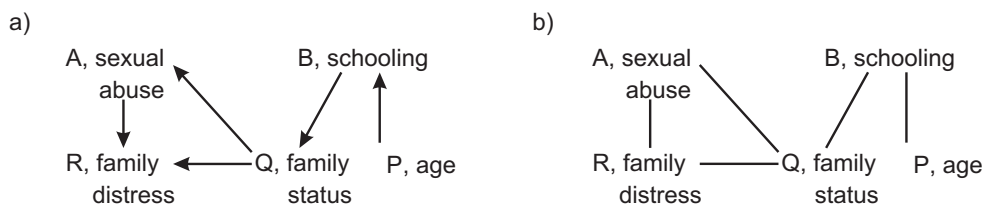


Figure 2: Two Markov equivalent graphs to the one of Y_b, Y_c of Figure 1.

From knowing the Markov equivalence to the graph in Figure 2a), the joint response model for Y_b given Y_a may also be fitted in terms of univariate regressions and from the Markov equivalence to the graph in Figure 2b), one knows for instance directly, using Proposition 1 below, that sexual abuse is independent of age and schooling given knowledge about family distress and family status.

Regression graphs are a subclass of *the maximal ancestral graphs* of Richardson and Spirtes (2002) and these are a subclass of the *summary graphs* of Wermuth (2011). The two types are called *corresponding graphs* if they result after marginalising over a node set m and conditioning on a disjoint node set c from a given directed acyclic graph. Both are *independence-preserving graphs* in the sense that they give the independence structure implied by the generating graph for the remaining nodes. The summary graph permits in addition to trace possible distortions of generating dependences as they arise in conditional associations among the remaining variables, for instance in parameters of the maximal ancestral graph models.

In the following Section 2, we introduce further concepts and the notation needed to state at the end of Section 2, some of the main results of the paper and related results

in the literature. In Section 3, a well-fitting regression graph is derived for data of chronic pain patients. Sections 4, 5 and 6 may be skipped if one wants to turn directly to formal definitions, new results and proofs in Section 7. Section 4 reviews linear recursion relations that are mimicked by graphs and lead to the standard and to special ways of combining probability statements, summarized here in Section 5. In Section 6, some of the previous results in the literature for graphs and for Markov equivalences are highlighted. The Appendix contains details of the regressions analyses in Section 3.

2 Some further concepts and notation

Figure 3 shows five ordered blocks, to introduce the notion of connected components of the graph to represent conditionally independent responses given their common past.

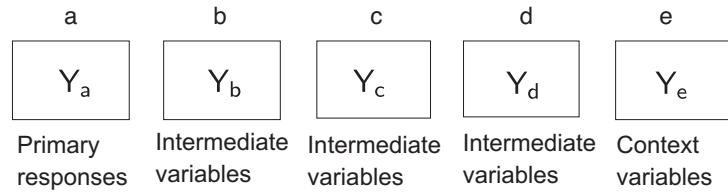


Figure 3: A typical first ordering: here of five vector variables, Y_a, \dots, Y_e ; primary response Y_a listed on the left, context variable Y_e on the right, intermediate variables in between.

In the example of a regression graph in Figure 4 corresponding to Figure 3, Y_a is a single response, Y_b has two component variables, both of Y_c and Y_e have four and Y_d has three. Each of the blocks b to e shows two **stacked boxes**, that is subsets of nodes without any undirected edge joining them. This indicates that disconnected components of a given vector variable are conditionally independent given their past.

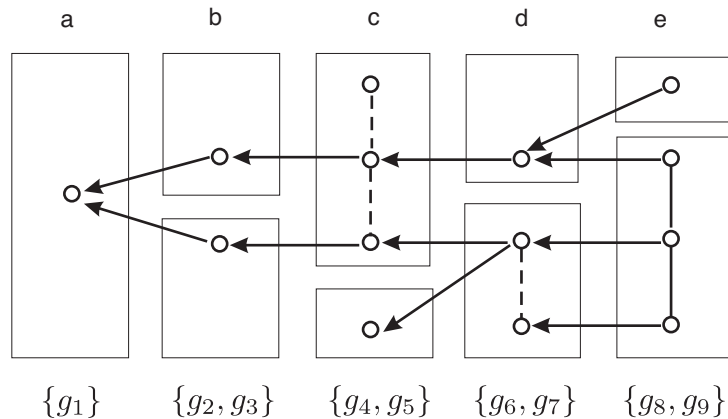


Figure 4: A regression graph for 14 variables corresponding to blocks a to e of Figure 3.

Graphs with dashed lines are *covariance graphs denoted by G_{cov}^N* , those with full lines are *concentration graphs denoted by G_{con}^N* ; see Wermuth and Cox (1998). The names are to remind one of their parametrisation in *regular joint Gaussian distributions*, in which the covariance matrix is invertible and gives the *concentration matrix*. A zero ik -element in G_{cov}^N means $i \perp\!\!\!\perp k$ and a zero ik -element in G_{con}^N means $i \perp\!\!\!\perp k | \{1, \dots, d\} \setminus \{i, k\}$; see Wermuth (1976a) or Cox and Wermuth (1996), Section 3.4.

The regression graph of Figure 4 is consistent with the first ordering in Figure 3 since there are no edges or only lines, i.e. undirected edges, within blocks a to e . After statistical analysis, blocks of the first ordering are often subdivided into the connected components of the graph, g_j , shown here in Figure 4 with the help of the stacked boxes. For several nodes in g_j , each pair of nodes is connected by at least one *path* within g_j , that is via a sequence of edges coupling distinct nodes.

For a regression graph, the *connected components g_j* , for $j = 1, \dots, J$, are the disconnected, undirected graphs that remain after removing all arrows from the graph. Thus, the displayed, stacked boxes in Figure 4 are just a visual aid. *Dashed lines* are for joint responses within g_1, \dots, g_r , say. *Full lines* are for context variables within g_{r+1}, \dots, g_J . The regression graph is *complete* if every node pair is coupled. In this case, the model is *saturated* as it is unconstrained for some given family of distributions.

For any one block of stacked boxes, different orderings are possible. If each arrow that starts at a node in any g_j points to a node in g_1, \dots, g_{j-1} , but never to a node in $g_{>j} = g_{j+1} \cup \dots \cup g_J$, then we speak of a *compatible ordering*.

Let g_1, \dots, g_J denote any compatible ordering of the connected components of G_{reg}^N , then a corresponding joint density factorises as

$$f_N = \prod_{j=1}^J f_{g_j | g_{>j}}. \quad (1)$$

The meaning of a missing edge for node pair i, k in G_{reg}^N is

- (i) $i \perp\!\!\!\perp k | g_{>j}$ for i, k both in g_j , $j = 1, \dots, r$,
- (ii) $i \perp\!\!\!\perp k | g_{>r} \setminus \{i, k\}$ for i, k both in $g_{>r}$,
- (iii) $i \perp\!\!\!\perp k | g_{>j} \setminus \{k\}$ for i in g_j with $j \leq r$ and k in $g_{>j}$.

Notice that only for context variables, conditioning is on variables within the same connected component while independence constraints on joint responses concern exclusively variables in their past.

Equation (2)(ii) holds for the concentration graph within $g_{>r}$ of the context variables having full lines as edges, equation (2)(i) for the conditional covariance graphs within g_j of joint responses having dashed lines as edges and (2)(iii) for the bipartite graph of dependences of the single responses within g_j on variables in the past of g_j having arrows

as edges. For instance in Figure 1, we have $S \perp\!\!\!\perp U|bc$ by (2)(i), $P \perp\!\!\!\perp Q|B$ by (2)(ii), and both $A \perp\!\!\!\perp B|PQ$ and $A \perp\!\!\!\perp P|BQ$ by (2)(iii).

Each missing edge of a regression graph corresponds to an independence statement for the uncoupled node pair; see equation (2), Lemma 2 and Lemma 3 below. Therefore, regression graphs represent one special class of the so-called *independence graphs*. Whenever a regression graph G_{reg}^N consists of *two disconnected graphs*, for Y_a and Y_b say, since no path leads from a node in a to a node in b , and $a \cup b = N$, then $a \perp\!\!\!\perp b$ or $f_N = f_a f_b$, and the two vector variables may be analysed separately. Therefore, we treat in Section 7 of this paper only connected regression graphs.

Graphs discussed in this paper *have no loops*, that is no edge connects a node to itself, and they have *at most one edge between two different nodes*. Recall that a path in such a graph can be described by a sequence of its nodes. The first and the last node in the sequence are the *endpoints of the path*, all other nodes are the distinct *inner nodes of the path*. An ik -path has endpoint nodes i and k . An ik -path without inner nodes is an edge. A path together with an edge between its endpoints is called a *cycle*.

A three-node path of arrows may contain only one of the three types of inner nodes shown in Figure 5, called *transition*, *source* and *sink node*, respectively.

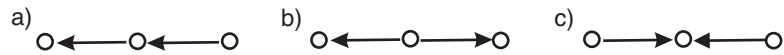


Figure 5: The three types of three-node paths in directed acyclic graphs with inner nodes named a) transition, b) source, c) sink node (or in directed acyclic graphs: collision node).

A *path is directed* if all its inner nodes are transition nodes. In a *directed cycle*, all edges are arrows pointing in the same direction and one returns to any starting node following the direction of the arrows. A regression graph contains no directed cycle but possibly *semi-directed cycles* which have at least one undirected edge in an otherwise directed cycle. If an arrow starts on a directed ik -path at k and points to i then node k has been named an *ancestor* of node i and node i a *descendant* of node k .

The *subgraph induced by a subset a* of the node set N consists of the nodes within a and of the edges present in the graph within a . A special type of induced subgraph, needed here, consisting of three nodes and two edges, is named a *V-configuration* or just a *V*. Thus, a three-node path forms a *V* if the induced subgraph has two edges.

A *path is chordless* if each of its three consecutive nodes that include two inner nodes, form a *V* in the graph. In a *chordless cycle* of four or more nodes, the subgraph induced by every consecutive three inner nodes is a *V*. An *undirected graph is chordal* if it contains no chordless cycle in four or more nodes.

In regression graphs, there may occur the three types of *collision Vs* shown in Figure 6. Notice that in a directed acyclic graph, the only possible collision V is directed and coincides with the sink V of Figure 5c).



Figure 6: The three types of collision V s in regression graphs: a) undirected, b) directed or sink-oriented, c) semi-directed.

An important common feature of the three V s of Figure 6 is that in G_{reg}^N , the inner node is excluded from every independence statements for the endpoints; see Lemma 2. In all other five possible types of V -configurations of a regression graph, named *transmitting Vs*, the inner node is instead included in the independence statement for the endpoints; see Lemma 3 below. Notice that for uncoupled endpoints, both paths a) and b) of Figure 5 are transmitting V s.

A *collision path* has as inner nodes exclusively collision nodes, while a *transmitting path* has as inner nodes exclusively transmitting nodes. A chordless collision path in four nodes contains at least one dashed line. In particular, it is impossible to replace all the edges in such a four-node path by arrows and not generate at least one transmitting V . Thereby, the meaning of this missing edge would then be changed. The *skeleton* of a graph results by replacing each edge present by a full line. Now, two of the main new results of this paper can be stated.

Theorem 1. *Two regression graphs are Markov equivalent if and only if they have the same skeleton and the same sets of collision V s, irrespective of the type of edge.*

Theorem 2. *A regression graph with a chordal graph for the context variables can be oriented to be Markov equivalent to a directed acyclic graph in the same skeleton, if and only if it does not contain any chordless collision path in four nodes.*

Sequences of regressions were introduced and studied, without specifying a concentration graph model for the context variables, by Cox and Wermuth (1993), Wermuth and Cox (2004), under the name of multivariate regression chains, reminding one of the sequences of unconstrained models that the class contains for Gaussian joint responses. An extension to graphs including a concentration graph had already been proposed for directed acyclic graph by Kiiveri, Speed and Carlin (1984). By this type of extension, the so-called *global Markov property of the graph* remains unchanged. This property permits to read off the graph all independence statements implied by the graph.

A criterion for Markov equivalence of summary graphs has been derived by Sadeghi (2009) who also shows that two different criteria for maximal ancestral graphs are equivalent, those due to Zhao, Zheng and Liu (2005) and to Ali, Richardson and Spirtes

(2009). These available Markov equivalence results and the associated proofs increase considerably in complexity, the larger the model class. On the other hand, the Markov equivalence criterion of Theorem 1 is simple and includes as special cases all available equivalence results for directed acyclic graphs, for covariance graphs and for concentration graphs, as set out in detail in Sections 6 and 7 here.

For context variables taken as given, Gaussian regression graph models coincide with a large subclass of structural equation models (SEMs), those permitting local modeling due to the factorisation property (1) and are without any *endogeneous responses*. Such responses have residuals that are correlated with some of its regressors. For traditional uses of SEMs see for instance Jöreskog (1981), Bollen (1989), Kline (2006), while Pearl (2009) advocates SEMs as a framework for causal inquiries. In the econometric literature thirty years ago, independences were always regarded as ‘overidentifying’ constraints.

For discrete variables, more attractive features of regression graph models were derived by Drton (2009), who speaks of chain graph models of type IV for multivariate regression chains. He proves that each member in this class belongs to a curved exponential family, for a discussion of this notion see for instance Cox (2006). Discrete type IV models form also a subclass of marginal models; see Rudas, Bergsma and Nemeth (2010), Bergsma and Rudas (2002). Defining local independence statements that involve only variables in the past are equivalent to more complex local independences used by Drton (2009); see Marchetti and Lupporelli (2010). These local definitions imply the pairwise independences of equation (2) for any regression graph, G_{reg}^N .

Two other types of chain graph have been studied as joint response models in statistics, the so-called *AMP chain graphs* of Andersson, Madigan and Perlman (2001), and the *LWF chain graphs* of Lauritzen and Wermuth (1989) and Frydenberg (1990). These are suitable for modeling data from intervention studies, when they are Markov equivalent to a regression graph, since they have in common that pairwise independences include other nodes of the same connected component. For AMP graphs, in equation (2) (i) is replaced by

$$(i') \quad i \perp\!\!\!\perp k | g_{>j-1} \setminus \{i, k\} \quad \text{for } i, k \text{ both in } g_j, \quad j = 1, \dots, r,$$

and for LWF graphs, (i) is also replaced by (i') and (iii) by

$$(iii') \quad i \perp\!\!\!\perp k | g_{>j-1} \setminus \{i, k\} \quad \text{for } i \text{ in } g_j \text{ with } j \leq r \text{ and } k \text{ in } g_{>j}.$$

Not yet systematically approached is the search for simple *covering models that capture most but not all independences*, but for regression graphs see Propositions 8 to 10. Often, a covering model may be easier to fit than a reduced edge-minimal model; see Cox and Wermuth (1990) and the discussion of Figures 16 and 17 in Section 7.

Before we discuss the meaning of different types of missing edges for linear models in more detail, we derive a well-fitting regression graph for data given by Kappesser (1997).

3 Deriving and interpreting a regression graph

For 201 chronic pain patients, the role of the site of pain during a three week stay in a chronic pain clinic was to be examined. In this study, it was of main interest to investigate the changes in two main symptoms and to understand determinants of the overall treatment success as rated by the patients, three months after they had left the clinic.

Figure 7 shows a first ordering of the variables derived in discussions between psychologists, physicians and statisticians, which shows only those variables that remained relevant after statistical analyses. The first ordering of the variables gives for each single or joint response a list of its possible explanatory variables, shown in boxes to the right. Selecting for each response all its directly explanatory variables from this list and checking for remaining associations among components of joint responses, provides enough insight to derive a well-fitting regression graph model. With this type of local modeling, the reasons for the model choice are made transparent; see the appendix.

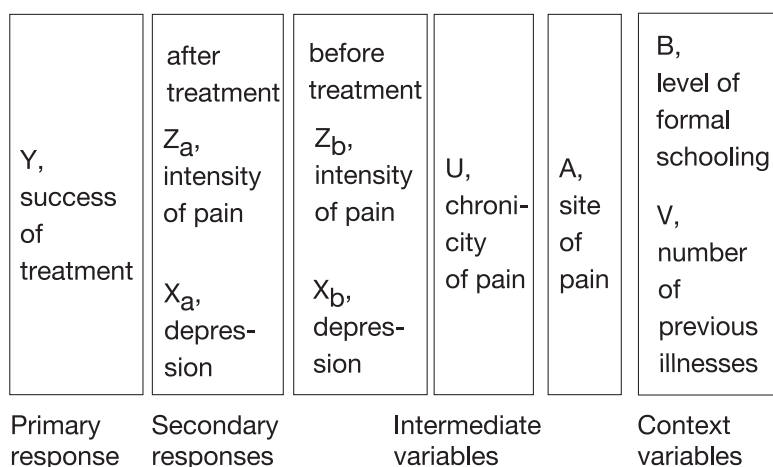


Figure 7: Initial ordering of variables in chronic pain study. There two joint responses, intensity of pain and depression They are the main symptoms of chronic pain measured here before and after treatment. The components of each joint response and of each single response are considered conditionally given the variables listed in boxes to their right.

Of the available background variables, age, gender, marital status and others, only the binary variables, level of formal schooling (1:=less than ten years, 2:= ten or more years) and the number of previous illnesses in years (min:=0, max:=16) are displayed

in the far right box as the relevant context variables. The response of primary interest, self-reported success of treatment, is listed in the box to the far left. It is a score that ranges between 0 and 35, combining a patient's answers to a specific questionnaire.

There are a number of intermediate variables. These are both explanatory for some variables and responses to others. Of these, two are regarded as joint responses since they represent two symptoms of a patient, intensity of pain and depression. Both are measured before treatment and directly after treatment, that is after the three-week stationary stay. Questionnaire scores are available of depression (min:=0, max:=46) and of the self-reported intensity of pain (min:=0, max:=10).

The chronicity of pain is a score (min:=0, max:=8) that incorporates different aspects, such as the frequency and duration of pain attacks, the spreading of pain, the use of pain relievers, the patient's pain treatment history. In this study, there are only patients included with one of two main sites of pain, the pain is either on their upper body, 'head, face, or neck' or on their 'back'.

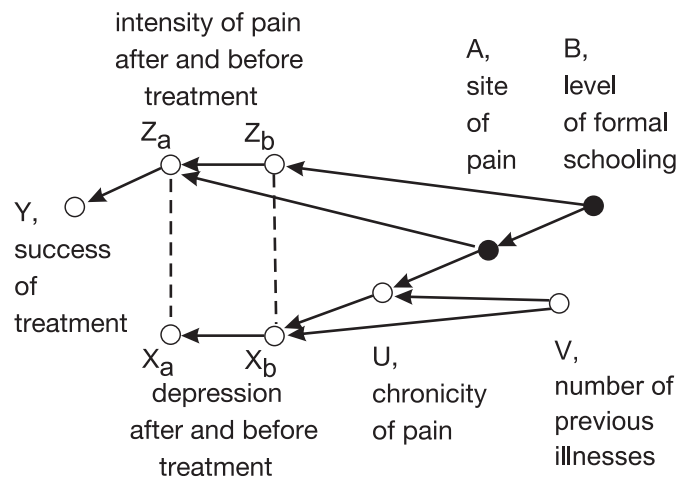


Figure 8: Regression graph, well compatible with the data, that results from the reported statistical analyses. Discrete variables are drawn as dots, continuous ones as circles.

The regression graph shown in Figure 8 summarizes some important aspects of the results of the statistical analyses for which more detail is given below. It tells in particular which of the variables are directly explanatory, that is which are important for predicting a response, by showing arrows that start from each of these directly explanatory variables and point to the response. Variables listed to the right of a response but without an arrow ending at this response do not substantially improve the prediction of the response when used in addition to the directly explanatory variables. For instance, for treatment success, only the pain intensity after the clinic stay is directly explanatory and this pain intensity is an important mediator (intermediate variable) between treatment success

and site of pain.

Scores of self-reported treatment success are low for almost all patients with high pain scores after treatment; see Figure 9. Otherwise, treatment success is typically judged to

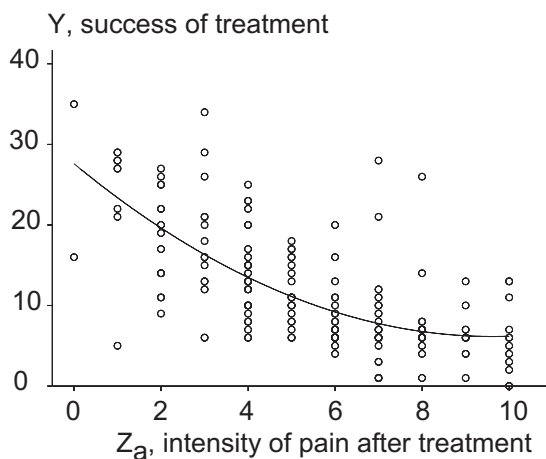


Figure 9: Form of dependence of primary response Y on Z_a .

be higher the lower the intensity of pain after treatment. This explains the nonlinear dependence of Y on Z_a .

As mentioned before, for back pain patients, the chronicity scores are on average higher than for head-ache patients and connected with a higher chronicity of the pain are higher scores of depression. These patients may have tried too late, after the acute pain had started, to get well focused help. Both before and after treatment, highly depressed patients tend also to report higher intensities of pain than others. The study provides no information which variables may explain these associations between the symptoms that remain after having taken their available explanatory variables into account. However, hidden common explanatory variables may exist in both cases since these remaining associations between the symptoms do not depend systematically on any other observed variable.

Some variables are indirectly explanatory. For each indirectly explanatory variable, a sequence of two or more arrows points to the response variable. For instance, the level of formal schooling and the site of pain are both indirectly explanatory for each of the symptoms after treatment and for the overall treatment success.

Once the types and directions of the direct dependence are taken into account, the regression graph helps to trace the development of chronic pain, starting from the context information on the level of schooling and the number of previous illnesses of a patient. Thus, patients with more years of formal schooling are more likely to be chronic head-ache patients. Patients with a lower level of formal schooling are more likely to be

back-ache patients, possibly because more of them have jobs involving hard physical work. Back-ache patients reach higher stages of the chronicity of pain and report higher intensity of pain still after treatment and are therefore typically less satisfied with the treatment they had received.

Graphical screening for nonlinear relations and interactive effects (Cox and Wermuth, 1994) pointed to the nonlinear dependence of treatment success on intensity of pain after treatment but to no other such relations. The regression graph model is said to fit the data well because for each single response separately, there is no indication that adding a further variable would improve prediction. The seemingly unrelated dependences of the symptoms after treatment on those before treatment agree so well with the observations that they differ also little from regressions computed separately, see the appropriate tables in the appendix.

Had there been no nonlinear relation and no categorical variables as responses, the overall model fit could also have been tested within the framework of structural equation models once the regression graph is available. This graph is derived here with the local modeling steps that use the first ordering of the variables, just in terms of univariate, multivariate and seemingly unrelated regressions. The regression graph provides a hypothesis which may be tested locally and/or globally in future studies that include the same set of nine variables. In this case, no variable selection strategy would be used or needed.

The available results for changes of the regression graph (Wermuth, 2011) that result after marginalising and conditioning provide a solid basis for comparing the results of any sequence of regressions with studies that contain the same set of core variables but which have some of the variables omitted or which consider subpopulations, defined by levels or level combinations of other variables. For instance for comparisons with the current study, the same chronicity score may not be recorded in another pain clinic or data may be available only for patients with pain in the upper body.

The main results of this empirical study are that site of pain needs to be taken into account also in future studies since it is an important mediator between the intrinsic characteristics of a patient, measured here by the given context variables, for both the overall treatment success and for the symptoms after treatment. For back-ache patients, the chronicity of pain and the depression score is higher than for the head-ache patients and the treatment is less successful since the intensity of pain remains high after the treatment in the clinic.

In the following section we give three-variable examples of a Gaussian joint response regression and of the three subclasses of regression graphs which have only one type of edge, of the covariance, the concentration and the directed acyclic graph.

4 Regressions, associations and recursive relations

For a quantitative response with linear dependences, the simple regression model dates back at least several centuries. The fitting of a least-squares regression line had been developed separately by Carl Friedrich Gauss (1777–1855), Adrien-Marie Legendre (1752–1833) and Robert Adrain (1775–1843). The method extends directly to models with several explanatory variables.

The most studied regression models are for joint Gaussian distributions. As we shall see, regression graphs mimic important features of these linear models but represent also relations in other distributions of continuous and discrete variables, which permit in particular nonlinear and interactive dependences. In a joint Gaussian distribution, let the mean-centered vector variable Y have dimension three, then we write the covariance matrix, Σ , and the concentration matrix Σ^{-1} , with graphs shown in Figure 10, as

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \cdot & \sigma_{22} & \sigma_{23} \\ \cdot & \cdot & \sigma_{33} \end{pmatrix}, \quad \Sigma^{-1} = \begin{pmatrix} \sigma^{11} & \sigma^{12} & \sigma^{13} \\ \cdot & \sigma^{22} & \sigma^{23} \\ \cdot & \cdot & \sigma^{33} \end{pmatrix},$$

where the dot-notation indicates entries in a symmetric matrix.



Figure 10: For unconstrained trivariate Gaussian distributions, the parameters attached to the edges are those corresponding to a) a covariance graph, b) a concentration graph.

With the edge of node pair $(1, 2)$ removed, both graphs turn into a V but have different interpretations. The resulting independence constraints are for Figures 10 a) and b), respectively,

$$1 \perp\!\!\!\perp 2 \iff (\sigma_{12} = 0) \quad \text{and} \quad 1 \perp\!\!\!\perp 2 | 3 \iff (\sigma^{12} = 0),$$

where the latter derives as an important property of concentration matrices; see Wermuth (1976a), Cox and Wermuth (1996), Section 3.4. For other distributions, the independence interpretation of these two types of undirected graph remains, but not the parametrisation. A similar statement holds for directed acyclic graphs and, more generally, for regression graphs, as illustrated here next.

For the linear equations that lead to a complete directed acyclic graph for a trivariate Gaussian distribution with mean zero, one starts with three mutually independent

Gaussian residuals ε_i and takes

$$\begin{aligned} Y_1 &= \beta_{1|2.3}Y_2 + \beta_{1|3.2}Y_3 + \varepsilon_1 \\ Y_2 &= \beta_{2|3}Y_3 + \varepsilon_2 \\ Y_3 &= \varepsilon_3. \end{aligned} \tag{3}$$

Because of this form, one speaks of triangular systems of equations also when the distribution of the residuals is not Gaussian, but the residuals are just uncorrelated. In that case, the probabilistic independence interpretation is lost, only the lack of linear relations can be inferred with a vanishing regression coefficient.

In econometrics, Hermann Wold (1908–1992) introduced such systems as linear recursive equations with uncorrelated residuals. Harald Cramér (1893–1985) used the term linear least-squares equations for residuals being uncorrelated with the regressors and the notation is an adaption of the one introduced by Udney Yule (1871–1951) and William Cochran (1909–1980) to indicate, for instance with $\beta_{1|3.2}$, that Y_1 is the response of regressor Y_3 and that the other regressor variable in the equation is Y_2 .

In joint Gaussian distributions, independence constraints on triangular systems mean vanishing equation parameters and missing edges in directed acyclic graphs, such as

$$1 \perp\!\!\!\perp 2|3 \iff (\beta_{1|2.3} = 0) \quad \text{and} \quad 2 \perp\!\!\!\perp 3 \iff (\beta_{2|3} = 0).$$

The complete directed acyclic graph corresponding to equations (3) is in Figure 11a)

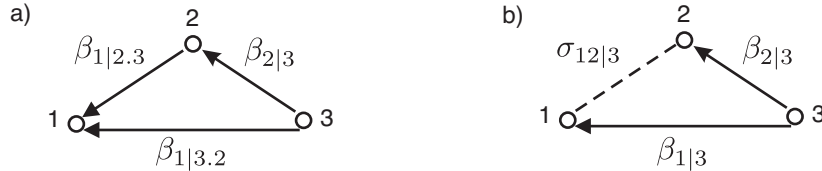


Figure 11: Parameters of a Gaussian distribution in: a) a complete G_{dag}^N , b) a complete G_{reg}^N .

For the smallest joint response model with the complete graph shown in Figure 11b), we take both Gaussian variables Y_1 and Y_2 to depend on a Gaussian variable Y_3 , to get equations (4) with residuals having zero means and being uncorrelated with Y_3 :

$$Y_1 = \beta_{1|3}Y_3 + u_1, \quad Y_2 = \beta_{2|3}Y_3 + u_2, \quad Y_3 = u_3. \tag{4}$$

Here, $\sigma_{12|3} = E(u_1u_2)$. The generating processes and hence the interpretation differs for the two models in equations (3) and (4). In the corresponding graphs of Figures 11a) and

11b), the vanishing of the edges for pairs (1,2) and (2,3) mean the same independence constraints since

$$1 \perp\!\!\!\perp 2|3 \iff (\sigma_{12|3} = 0) \iff (\beta_{1|2.3} = 0) \quad \text{and} \quad 2 \perp\!\!\!\perp 3 \iff (\beta_{2|3} = 0),$$

but the edges for pair (1,3) capture different conditional dependences.

Again, taking away any arrow generates a \mathbf{V} . Taking away any two edges means to combine two independence statements. This is discussed further in the next section.

One of the special important features of the linear least-squares regressions is that the residuals are uncorrelated with the regressors. The effect is that the model part coincides with a conditional linear expectation as illustrated here with a model for response Y_1 and regressors Y_2, Y_3 , which we take, as mentioned before, as measured in deviations from their means. For instance, one gets for

$$\begin{aligned} Y_1 &= \beta_{1|2.3}Y_2 + \beta_{1|3.2}Y_3 + \varepsilon_1, \\ E_{\text{lin}}(Y_1|Y_2, Y_3) &= \beta_{1|2.3}Y_2 + \beta_{1|3.2}Y_3. \end{aligned} \tag{5}$$

There is a recursive relation for least-squares regression coefficients; see Cochran (1938), Cox and Wermuth (2003), Ma, Xie and Geng (2006). It shows for instance with

$$\beta_{1|3} = \beta_{1|3.2} + \beta_{1|2.3}\beta_{2|3} \tag{6}$$

that $\beta_{1|3.2}$, the partial coefficient of Y_3 given also Y_2 as a regressor for Y_1 , coincides with the marginal coefficient, $\beta_{1|3}$, if and only if $\beta_{1|2.3} = 0$ or $\beta_{2|3} = 0$.

The method of maximizing the likelihood was recommended by Sir Ronald Fisher (1890–1962) as a general estimation technique that applies also to regressions with categorical or quantitative responses. One of the most attractive features of the method concerns properties of the estimates. Given two models with parameters that are in one-to-one correspondence, the same one-to-one transformation leads from the maximum-likelihood estimates under one model to those of the other.

Different single response regressions, such as logistic, probit, or linear regressions, were described as special cases of the generalized linear model by Nelder and Wedderburn (1972); see also McCullagh and Nelder (1989). In all of these regressions, the vanishing of the coefficient(s) of regressors indicates conditional independence from the response given all remaining regressors in the regression model.

The general linear model with a vector response, also called multivariate linear regression, has identical sets of regressors for each component of the vector and the individual component variables of the response vector form the set of joint responses. Maximum-likelihood estimation of regression coefficients for a joint Gaussian distribution reduces to linear-least squares fitting for each component separately; see Anderson (1958).

With different sets of regressors for the components of a vector response, seemingly unrelated regressions (SUR) result and iterative methods are needed for estimation; see Zellner (1962). For small sample sizes, a given solution of the likelihood equations of a Gaussian SUR model may not be unique; see Drton and Richardson (2004), Sundberg (2010), while for exclusively discrete variables this will never happen; see Drton (2009).

For mixed variables, no corresponding results are available yet. But in general, there often exists a covering model with nice estimation properties. For instance for the SUR model with regression graph

$$\circ \longrightarrow \circ \text{ --- } \circ \longleftarrow \circ ,$$

a general linear model with two independent regressors is a simple covering model.

For a vector variable of categorical responses only, the multivariate logistic regression of Glonek and McCullagh (1995) reduces to separate main effect logistic regressions for each component of the response vector provided that certain higher-order interactions vanish; see Marchetti and Lupporelli (2010). In the context of structural equation models (SEMs), dependences of binary categorical variables are modeled in terms of probit regressions. These do not differ substantially from logistic regressions whenever the smallest and largest events occur at least with probability 0.1; see Cox (1966).

Multivariate linear regressions as well as SUR models belong to the framework of SEMs even though this general class had been developed in econometrics to deal with endogenous responses, defined by the existence of correlations between the residuals and some regressors. For endogenous responses, the equation parameters are no longer measures of conditional dependence, as they are in linear least-squares regression models. Estimation methods for SEMs were discussed in the Berkeley symposia on mathematical statistics and probability from 1945 to 1965, but some issues still need to be settled; see for instance Drton, Eichler and Richardson (2009), Stanghellini and Wermuth (2005).

In statistical models that treat all variables on equal standing, the variables are not assigned roles of responses or regressors and undirected measures of association are used instead of coefficients of dependence. In the concentration graph models, the undirected associations are conditional given all remaining variables on equal standing.

For instance, for categorical variables, these models are better known as graphical log-linear models; see Birch (1963), Caussinus (1966), Goodman (1970), Bishop, Fienberg and Holland (1975), Wermuth (1976a), Darroch, Lauritzen and Speed (1980). For Gaussian random variables, these had been introduced as covariance selection models; see Dempster (1972), Wermuth (1976b), Speed and Kiiveri (1986), Kiiveri (1987), Drton and Perlman (2004), and for mixed variables as graphical models for conditional Gaussian (CG) distributions; see Lauritzen and Wermuth (1989), Edwards (2000).

For a mean-centered vector variable Y , the elements of the covariance matrix Σ are

$\sigma_{ij} = E(Y_i Y_j)$. If Σ is invertible, the covariances σ_{ij} are in a one-to-one relation with the concentrations σ^{ij} , the elements of the concentration matrix Σ^{-1} . There is a recursive relation for concentrations; see Dempster (1969). For a trivariate distribution

$$\sigma^{23.1} = \sigma^{23} - \sigma^{12}\sigma^{13}/\sigma^{11}, \quad (7)$$

where $\sigma^{23.1}$ denotes the concentration of Y_2, Y_3 in their bivariate marginal distribution. Thus, the overall concentration σ^{23} coincides with $\sigma^{23.1}$ if and only if $\sigma^{12} = 0$ or $\sigma^{13} = 0$.

Alternatively in covariance graph models, the undirected measures for variables on equal standing are pairwise marginal associations. For Gaussian variables, these models had been introduced as hypotheses linear in covariances; see Anderson (1973), Kauermann (1996), Wermuth, Cox and Marchetti (2006), Chaudhuri, Drton and Richardson (2007). For categorical variables, covariance graph models have been studied only more recently; see Drton and Richardson (2008), Lupporelli, Marchetti and Bergsma (2009). Again, no similar estimation results are available for general mixed variables yet.

There is also a recursive relation for covariances; see Anderson (1958). It shows for instance, for just three components of Y having a Gaussian distribution, with

$$\sigma_{12|3} = \sigma_{12} - \sigma_{13}\sigma_{23}/\sigma_{33}, \quad (8)$$

where $\sigma_{12|3}$ denotes the covariance of Y_1, Y_2 given Y_3 . Therefore, $\sigma_{12|3}$ coincides with σ_{12} if and only if $\sigma_{13} = 0$ or $\sigma_{23} = 0$. By equations (6), (7), (8), a unique independence statement is associated with the endpoints of any V in a trivariate Gaussian distribution.

In the context of multivariate exponential families of distributions, concentrations are special canonical parameters and covariances are special moment parameters with estimates of canonical and moment parameters being asymptotically independent; see Barndorff-Nielsen (1978). Regression graphs capture independence structures for more general types of distribution, where operators for transforming graphs mimic operators for transforming different parametrisations of joint Gaussian distributions; see Wermuth, Wiedenbeck and Cox (2006), Wiedenbeck and Wermuth (2010), Wermuth (2011).

In particular, by removing an edge from any V of a regression graph, one introduces an additional independence constraint just as in a regular joint Gaussian distribution. This requires that the generated distributions satisfy the composition and intersection property in addition to general properties, as discussed in the next section.

5 Using graphs to combine independence statements

We now state the four standard properties of independences of any multivariate distribution; see e.g. Dawid (1979), Studený (2005), as well as two special properties of joint

Gaussian distributions. The six taken together, describe the combination and decomposition of independences in regression graphs, for instance those resulting by removing edges. We discuss when these six properties apply also to regression graph models.

Let X, Y, Z be random (vector) variables, continuous, discrete or mixed. By using the same compact notation, f_{XYZ} for a given joint density, a probability distribution or a mixture and by denoting the union of say X and Y by XY , one has

$$X \perp\!\!\!\perp Y|Z \iff (f_{XYZ} = f_{XZ}f_{YZ}/f_Z), \quad (9)$$

where for instance f_Z denotes the marginal density or probability distribution of Z . Since the order of listing variables for a given density is irrelevant, ***symmetry of conditional independence*** is one of the standard properties, that is

$$(i) \quad X \perp\!\!\!\perp Y|Z \iff Y \perp\!\!\!\perp X|Z.$$

Equation (9) restated for instance for the conditional distribution of X given Y and Z , $f_{X|YZ} = f_{XYZ}/f_{YZ}$, is

$$X \perp\!\!\!\perp Y|Z \iff (f_{X|YZ} = f_{X|Z}). \quad (10)$$

When two edges are removed from a graph in Figures 10 and 11, just one coupled pair remains, suggesting that the single node is independent of the pair.

For instance in Figure 11a), with nodes 1, 2, 3 corresponding in this order to X, Y, Z , removing the arrows for (1,2) and (2,3), leaves (1,3) disconnected from node 2. For any joint density, implicitly generated as $f_{XYZ} = f_{X|YZ}f_{Y|Z}f_Z$, one has equivalently,

$$(X \perp\!\!\!\perp Y|Z \text{ and } Y \perp\!\!\!\perp Z) \iff XZ \perp\!\!\!\perp Y.$$

In general, the ***contraction property*** is for a, b, c, d disjoint subsets of N :

$$(ii) \quad (a \perp\!\!\!\perp b|cd \text{ and } b \perp\!\!\!\perp c|d) \iff ac \perp\!\!\!\perp b|d.$$

It has become common to say that a distribution is generated over a given G_{dag}^N if the variables are generated in any order compatible with the graph and the distribution factorizes as specified by the graph. For instance, for a trivariate distribution generated over the collision \mathbf{V} of Figure 11b) obtained by removing the edge for (2,3), both orders (1, 2, 3) and (1, 3, 2) are compatible with the graph and $f_{XYZ} = f_{X|YZ}f_Yf_Z$.

Conversely, suppose that $XZ \perp\!\!\!\perp Y$ holds, then this implies $X \perp\!\!\!\perp Y$ and $Z \perp\!\!\!\perp Y$ so that for instance the same two edges as in Figure 11b) are missing in the corresponding covariance graph of Figure 10a). In general, the ***decomposition property*** is for a, b, c, d disjoint subsets of N :

$$(iii) \quad a \perp\!\!\!\perp bc|d \implies (a \perp\!\!\!\perp b|d \text{ and } a \perp\!\!\!\perp c|d).$$

In addition, $XZ \perp\!\!\!\perp Y$ implies $X \perp\!\!\!\perp Y|Z$ and $Z \perp\!\!\!\perp Y|X$ so that for instance the same two edges as in Figure 11a) are missing in the corresponding concentration graph of Figure 10b). In general, the **weak union property** is for a, b, c, d disjoint subsets of N :

$$(iv) a \perp\!\!\!\perp bc|d \implies (a \perp\!\!\!\perp b|cd \text{ and } a \perp\!\!\!\perp c|bd).$$

Under some regularity conditions, all joint distributions with a given density f_N share the four properties (i) to (iv).

Joint distributions, for which the reverse implication of the decomposition property (iii) and of the weak union property (iv) hold such as a regular joint Gaussian distribution, are said to have, respectively, the **composition property** (v) and the **intersection property** (vi), that is for a, b, c, d disjoint subsets of N :

$$(v) (a \perp\!\!\!\perp b|d \text{ and } a \perp\!\!\!\perp c|d) \implies a \perp\!\!\!\perp bc|d,$$

$$(vi) (a \perp\!\!\!\perp b|cd \text{ and } a \perp\!\!\!\perp c|bd) \implies a \perp\!\!\!\perp bc|d.$$

The standard graph theoretical separation criterion has different consequences for the two types of undirected graph corresponding for Gaussian distributions to concentration and to covariance matrices. We say **a path intersects subset set c** of node set N if it has an inner node in c and let $\{a, b, c, m\}$ partition N to formulate known Markov properties. The notation is to remind one that with any independence statement $a \perp\!\!\!\perp b|c$, one implicitly has marginalised over the remaining nodes in $m = V \setminus \{a \cup b \cup c\}$, i.e. one considers the marginal joint distribution of Y_a, Y_b, Y_c .

Proposition 1. Lauritzen (1996). *A concentration graph, G_{con}^N , implies $a \perp\!\!\!\perp b|c$ if and only if every path from a to b intersects c .*

Proposition 2. Kauermann (1996). *A covariance graph, G_{cov}^N , implies $a \perp\!\!\!\perp b|c$ if and only if every path from a to b intersects m .*

Notice that Proposition 1 requires the intersection property and Proposition 2 requires the composition property.

A subgraph induced by nodes $a \cup b$ in G_{cov}^N is the covariance graph $G_{\text{cov}}^{a \cup b}$. We say that there is **an edge between a and b** if there is an edge with one node in a and the other node in b . In this case, the graph $G_{\text{cov}}^{a \cup b}$ **is connected in a and b**, otherwise **the graphs in a and b are disconnected**.

Corollary 1. *A covariance graph, G_{cov}^N , or a concentration graph, G_{con}^N , implies $a \perp\!\!\!\perp b$ if and only if in the subgraph induced by $a \cup b$, the graphs in a and b are disconnected.*

It can be shown that the independence structure of a regression graph is fully specified by the pairwise independences (2) of each missing edge if both properties (v) and (vi) hold in addition to the standard ones; see also Kang and Tian (2009), Pearl and Paz (1987), Marchetti and Lupparelli (2010) for relevant, previous special results.

Lemma 1. *A regression graph, G_{reg}^N , captures an independence structure for a distribution with density f_N factorizing as (1) if properties (i) to (vi) hold for f_N .*

Proof. The first four properties hold for any density f_N . Given the intersection property (vi), any node i with missing edges to nodes k, l in a concentration graph of node set N implies $i \perp\!\!\!\perp \{k, l\} | N \setminus \{i, k, l\}$ and given the composition property (v), any node i with missing edges to nodes k, l in a covariance graph given Y_c implies $i \perp\!\!\!\perp \{k, l\} | c$. \square

For purely discrete and for Gaussian distributions, necessary and sufficient conditions for the intersection property (vi) to hold are known; see San Martin, Mouchart and Rolin (2005). Too strong sufficient conditions are for joint Gaussian distributions that they are regular and for discrete variables, that the probabilities are strictly positive.

The composition property (v) is known to hold for Gaussian distributions and for triangular binary distributions with at most main effects in symmetric $(-1, 1)$ variables; see Wermuth, Marchetti and Cox (2009).

Both properties (v) and (vi) hold, whenever a distribution may have been generated over a so-called parent graph; see Wermuth (2011), Marchetti and Wermuth (2009), Wermuth, Wiedenbeck and Cox (2006). Parent graphs, denoted by G_{par}^N , are directed acyclic graphs with some added properties.

Parent graphs are connected directed acyclic graphs to which one fixed, compatible ordering of the nodes is attached and the graph is edge-minimal. An **ordering of the nodes is compatible with a given G_{dag}^N** if each ancestor of any node i is within $g_{>i} = \{i + 1, \dots, d\}$. In addition, **G_{par}^N is to be edge-minimal for the generated distribution**, that is no edge can be removed from the graph without adding another independence statement to the distribution generated over the graph.

Also the process of generating distributions over G_{par}^N has special features. The parameters of $f_{i|>i}$ are taken to be variation-independent of those of $f_{>i}$ to assure maximization of the joint likelihood by separate fitting of each univariate regression. Consequences are that distributions generated over G_{par}^N , every collision \mathbf{V} is association-inducing for its endpoints by conditioning on the inner node and every transmitting \mathbf{V} is association-inducing by marginalising over the inner node, just like in a regular joint Gaussian distribution, see as examples the recursive relations of (6), (7), (8).

With distributions generated over parent graphs, one excludes incomplete families of distributions; see Lehmann and Scheffé (1955), Brown (1986), Mandelbaum and

Rüschendorf (1987), in which independence statements connected with a V may have the inner node both within and outside the conditioning set; see Wermuth (2011), Wermuth and Cox (2004), Darroch (1962). Such independences have been characterized as being not representable in joint Gaussian distributions; see Lněnička and Matúš (2007).

More generally, such independences cannot occur whenever the distribution is **weakly transitive** that is if, for i, k, l distinct nodes of N and $m = N \setminus \{i, k, l\}$,

$$(i \perp\!\!\!\perp k|l \text{ and } i \perp\!\!\!\perp k|\{l, m\}) \implies (i \perp\!\!\!\perp m|l \text{ or } k \perp\!\!\!\perp m|l).$$

or for $a \text{ dep } b|c$ denoting dependence of Y_a on Y_b given Y_c for disjoint subsets a, b, c of V and an edge-minimal graph, equivalently,

$$(i \text{ dep } m|l \text{ and } k \text{ dep } m|l \text{ and } i \perp\!\!\!\perp k|l) \implies i \text{ dep } k|\{l, m\}$$

and

$$(i \text{ dep } m|l \text{ and } k \text{ dep } m|l \text{ and } i \perp\!\!\!\perp k|\{l, m\}) \implies i \text{ dep } k|l.$$

Thus, there is in particular a unique independence interpretation associated with each V . For three categorical variables, an example of a distribution that is not weakly transitive has been given with equation (5.4) by Birch (1963) and, for jointly Gaussian variables and two nodes within m , with equation (8) by Cox and Wermuth (1993).

Special types of **distribution are said to be faithful to a graph** if every independence statement is captured by a given independence graph; see Spirtes, Glymour and Scheines (1993). This requires that (1) the graph is edge-minimal, (2) the distribution is weakly transitive, and (3) associations introduced by several active paths for any pair (i, k) do not cancel. An active path is specified here for regression graphs in Definition 1 and for directed acyclic graphs, it has been called d-connecting in the literature.

This faithfulness property imposes peculiar constraints on parameters whenever more than two nodes induce a complete subgraph in the graph; see for instance Figure 1 in Wermuth, Marchetti and Cox (2009), but it reduces to edge-minimality and weak transitivity whenever the graph is a **tree**, that is a connected graph with one unique path joining any node pair or it is a **forest**, that is a union of disjoint trees.

Every regression graph can be generated by a larger directed acyclic graph; see Richardson and Spirtes (2002). For regression graph models however, necessary and sufficient conditions are not yet known to assure that a joint distribution could have resulted from a larger parent graph model. One set of sufficient conditions, satisfied in the examples of this paper, uses chordal graphs and chordless paths. A regression graph model may result from a parent graph model in more nodes if (1) the subgraphs induced by connected response nodes form chordless paths, (2) the concentration graph of the context variables is chordal and (3) the association of each response variable pair can

be generated by some unobserved variable. For discrete variables, the last condition is always satisfied whenever an association does not depend systematically on the values of a third variable; see also Holland and Rosenbaum (1986).

6 Some early results on graphs and Markov equivalence

In the past, results concerning graphs and Markov equivalence have been obtained quite independently in the mathematical literature on characterizing different types of graph, in the statistical literature on specifying types of multivariate statistical models, and in the computer science literature on deciding on special properties of a given graph or on designing fast algorithms for transforming graphs.

For instance, following the simple enumeration result for *labeled trees* in d nodes, d^{d-2} , by Karl-Wilhelm Borchardt (1817-1880), it could be shown that these trees are in one-to-one correspondence to distinct strings of size $d - 2$; see Cayley (1889). Much later, labeled trees were recognized to form the subclass of directed acyclic graphs with exclusively source Vs and therefore to be also Markov equivalent to chordal concentration graphs; see Castelo and Siebes (2003).

In the literature on graphical Markov models, a number of different names have been in use for a sink V, for instance ‘two arrows meeting head-on’ by Pearl (1988), ‘unshielded collider’ by Richardson and Spirtes (2002), and ‘Wermuth-configuration’ by Whittaker (1990), after it had been recognized that, for Gaussian distributions, the parameters of a directed acyclic graph model without sink Vs are in one-to-one correspondence to the parameters in its skeleton concentration graph model.

Proposition 3. (Wermuth, 1980), (Wermuth and Lauritzen, 1983), (Frydenberg, 1990). *A directed acyclic graph is Markov equivalent to a concentration graph of the same skeleton if and only if it has no collision V.*

Efficient algorithms to decide whether an undirected graph can be oriented into a directed acyclic graph, became available in the computer science literature under the name of perfect elimination schemes; see Tarjan and Yannakakis (1984). When algorithms were designed later to decide which arrows may be flipped in a given G_{dag}^N , keeping the same skeleton and the same set of sink Vs, to get to a list of all Markov equivalent G_{dag}^N s, these results appear to have not been used; see Chickering (1995).

The number of equivalent characterizations of concentration graphs that have perfect elimination schemes has increased steadily, since they were introduced as rigid circuit graphs by Dirac (1961). These graphs without chordless cycles in four or more nodes are named also ‘chordal graphs’, ‘triangulated graphs’, ‘graphs with the running intersection property’ or ‘graphs with only complete prime graph separators.’

By contrast, for a covariance graph that can be oriented to be Markov equivalent to a G_{dag}^N of the same skeleton, chordless paths are relevant.

Proposition 4. (Pearl and Wermuth, 1994). *A covariance graph with a chordless path in four nodes is not Markov equivalent to a directed acyclic graph in the same node set.*

For distributions generated over directed acyclic graphs, sink Vs are needed again.

Proposition 5. (Frydenberg, 1990), (Verma and Pearl, 1990). *Directed acyclic graphs of the same skeleton are Markov equivalent if and only if they have the same sink Vs.*

Markov equivalence of a concentration graph and a covariance graph model is for regular joint Gaussian distributions equivalent to **parameter equivalence** which means that there is a one-to-one relation between the two sets parameters. Therefore, an early result on parameter equivalence for joint Gaussian distributions implies the following Markov equivalence result for distributions satisfying both the composition and the intersection property.

Proposition 6. (Jensen, 1988), (Drton and Richardson, 2008). *A covariance graph is Markov equivalent to a concentration graph if and only if both consist of the same complete, disconnected subgraphs.*

Fast ways of inserting an edge for every transition V , of deciding on connectivity and on blocking flows have been available in the corresponding Russian literature since 1970; see Dinitz (2006), but these results appear to have not been exploited for the so-called lattice conditional independence models, recognized as distributions generated over G_{dag}^N s without any transition Vs by Andersson, Madigan, Perlman and Triggs (1997).

Markov equivalence of other than multivariate regression chain graphs, have been given by Roverato (2005), Andersson and Perlman (2006) and Roverato and Studený (2006).

With the so-called global Markov property of a graph in node set N and any disjoint subsets a, b, c of N , one can decide whether the graph implies $a \perp\!\!\!\perp b | c$. To give this property for a regression graph, we use special types of path that has been called active; see Wermuth (2011). For this, let again $\{a, b, c, m\}$ partition the node set N of G_{reg}^N .

Definition 1. *A path from a to b in G_{reg}^N is active given c if its inner collision nodes are in c or have a descendant in c and its inner transmitting nodes are in $m = N \setminus (a \cup b \cup c)$. Otherwise, the path is said to break given c or to break with m .*

Thus, a path breaks when c includes an inner transmitting node or when m includes an inner collision node and all its descendants; see also Figure 4 of Marchetti and Wermuth (2009).

For directed acyclic graphs, an active path of Definition 1 reduces to the d-connecting path of Geiger, Verma and Pearl (1990). Similarly, the following proposition coincides in that special case with those of their so-called d-separation. Let node set N of G_{reg}^N be partitioned as above by $\{a, b, c, m\}$.

Proposition 7. (Cox and Wermuth, 1996), (Sadeghi, 2009). *A regression graph, G_{reg}^N , implies $a \perp\!\!\!\perp b|c$ if and only if every path between a and b breaks given c .*

Thus, whenever G_{reg}^N implies $a \perp\!\!\!\perp b|c$, this independence statement holds in the corresponding sequence of regressions for which the density f_N factorizes as (1), provided that f_N satisfies the same properties of independences, (i) to (vi) of Section 5, just like a regular Gaussian joint density. For example, in the graphs of Figure 12, node 2 is an ancestor of node 1 so that G_{reg}^N does not imply $3 \perp\!\!\!\perp 4|2$.

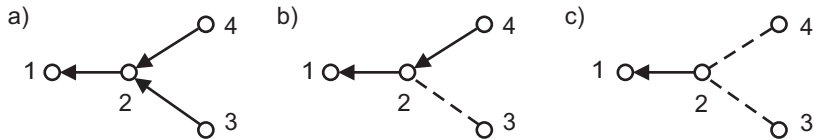


Figure 12: Three regression graphs which imply $3 \perp\!\!\!\perp 4$ but not $3 \perp\!\!\!\perp 4|2$.

Since covariance and concentration graphs consist only of one type of edge, the restricted versions in Propositions 1 and 2 of the defined path can be used for their global Markov property.

7 New results and proofs

We now treat connected regression graphs in node set N and corresponding distributions defined by sequences of regressions with joint discrete or continuous responses, ordered in connected components g_1, \dots, g_r of the graph, and with context variables in connected components, g_{r+1}, \dots, g_J , which factorize as in (1), satisfy the pairwise independences of (2) as well as properties of independence statements, given as (i) to (vi) in Section 5.

For the main result of Markov equivalence for regression graphs, we consider distinct nodes i and k , node subsets c of $N \setminus \{i, k\}$ and the notion of shortest active paths.

Definition 2. *An ik -path in G_{reg}^N is a shortest active path π with respect to c if every ik -path of G_{reg}^N with fewer inner nodes breaks given c .*

Every chordless π is such a shortest path. If the consecutive nodes (k_{n-1}, k_n, k_{n+1}) on $\pi = (i = k_0, k_1, \dots, k_m = k)$ induce a complete subgraph in G_{reg}^N , we say that there

is *a triangle on the path*. In Figure 13a) nodes 2,3,4 form a triangle on the path (1, 2, 4, 3, 5).

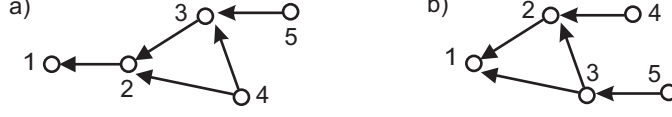


Figure 13: Graphs of active five-node paths a) with path (1, 2, 4, 3, 5) the shortest active path, where 3 is in c , b) active path (4, 2, 1, 3, 5), where 1 is in c , and a shorter active path (4, 2, 3, 5).

If this path is an active path connecting the uncoupled node pair (1,5), then nodes 2 and 4 are inner transmitting nodes outside c and the inner collision node 3 is in c . This path is then also the shortest active path connecting (1,5). The shorter path (1, 2, 3, 5) has nodes 2 and 3 as inner transmitting nodes, but is inactive since node 3 is in c .

By contrast in Figure 13b), when path (4, 2, 1, 3, 5) is an active path connecting the uncoupled node pair (4,5), then path (4,2,3,5) is a shorter active path. To see this, notice that on an active (4, 2, 1, 3, 5) path, the inner collision node 1 is in c and the inner transmitting nodes 2 and 3 are outside c . In this case, the inner collision node 2 on the path (4, 2, 3, 5) has node 1 as a descendant in c , so that this shorter path is also active.

We also use the following results for proving Theorem 1. The first two are direct consequences of Proposition 7 and imply the pairwise independences of equation (2). Lemma 4 results with (2). Let h, i, k be distinct nodes of N .

Lemma 2. For (h, i, k) a collision V in G_{reg}^N , the inner node i is excluded from c in every independence statement for h, k implied by G_{reg}^N .

Lemma 3. For (h, i, k) a transmitting V in G_{reg}^N , the inner node i is included in c in every independence statement for h, k implied by G_{reg}^N .

Lemma 4. A missing ik -edge in G_{reg}^N implies at least one independence statement $i \perp\!\!\!\perp k | c$ for c a subset of $N \setminus \{i, k\}$.

We can now derive the first of the main new results in this paper.

Theorem 1. Two regression graphs are Markov equivalent if and only if they have the same skeleton and the same sets of collision V s, irrespective of the type of edge.

Proof. Regression graphs G_{reg1}^N and G_{reg2}^N are Markov equivalent if and only if for every disjoint subsets a, b , and c of the node set of N , where only c can be empty,

$$(G_{\text{reg1}}^N \implies a \perp\!\!\!\perp b | c) \iff (G_{\text{reg2}}^N \implies a \perp\!\!\!\perp b | c). \quad (11)$$

Suppose first that (11) holds. By Lemma 4, G_{reg1}^N and G_{reg2}^N have the same skeleton, and by Lemma 2 and Lemma 3, G_{reg1}^N and G_{reg2}^N have the same collision Vs.

Suppose next that G_{reg1}^N and G_{reg2}^N have the same skeleton and the same collision Vs and consider two arbitrary distinct nodes i and k and any node subset c of $N \setminus \{i, k\}$. By Proposition 7, (11) is equivalent to stating that for every uncoupled node pair i, k , there is an active path with respect to c in G_{reg1}^N if and only if there is an active ik -path with respect to c in G_{reg2}^N .

Suppose further that path π is in G_{reg1}^N a shortest active ik -path with respect to c . Since G_{reg1}^N and G_{reg2}^N have the same skeleton, the path π exists in G_{reg2}^N . We need to show that it is active. If all consecutive two-edge-subpaths of π are Vs then π is active in G_{reg2}^N . Therefore, suppose that nodes (k_{n-1}, k_n, k_{n+1}) on π form a triangle instead of a V. It may be checked first, that in all other possible triangles in regression graphs that can appear on π than the two of Figure 14, there is as in Figure 13b) a shorter active path. To complete the proof, we show that for the two types of triangles shown in Figure 14a) and Figure 14b) path π is also in G_{reg2}^N an active ik -path with respect to c .



Figure 14: The two types of triangles in regression graphs without a shorter active path whenever the path with inner nodes (k_{n+1}, k_n, k_{n-1}) is active.

In G_{reg1}^N containing the triangle of Figure 14a) on a shortest active path π , node k_n is a transmitting node which is by Lemma 2 outside c . By Lemma 3, node k_{n-1} is a collision node inside c . If instead k_{n-1} were a transmitting node on π in G_{reg1}^N , it would also be a transmitting node on $(k_{n-2}, k_{n-1}, k_{n+1})$ and give a shorter active path via the $k_{n-1}k_{n+1}$ -edge, contradicting the assumption of π being a shortest path. Similarly, if collision node k_{n-1} on π were only an ancestor of c , then there were a shorter active path via the $k_{n-1}k_{n+1}$ -edge.

In addition, node pair k_n, k_{n-2} is uncoupled in G_{reg1}^N since by inserting any such edge, that is permissible in a regression graph, another shortest path via the $k_{n-2}k_n$ -edge would result. Therefore, since G_{reg1}^N and G_{reg2}^N have the same collision Vs, the subpath (k_{n-2}, k_{n-1}, k_n) forms also a collision V in G_{reg2}^N . Similarly, $(k_{n-2}, k_{n-1}, k_{n+1})$ is a transmitting V and (k_{n+2}, k_{n+1}, k_n) is a V of either type. Hence k_{n-1} is a parent of k_{n+1} in G_{reg2}^N and the only permissible edge between k_n and k_{n+1} is an arrow pointing to k_{n+1} . Therefore, π forms an active path also in G_{reg2}^N .

The proof for Figure 14b) is the same as for Figure 14a) since the type of nodes along π , i.e. as collision or transmitting nodes, are unchanged. \square

In the example of Figure 15, all three regression graphs have the same skeleton.

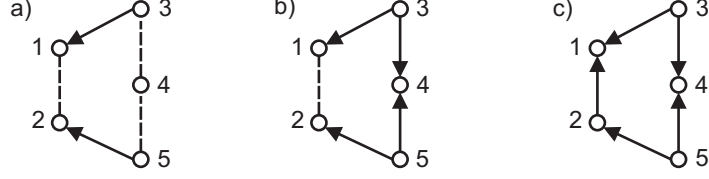


Figure 15: a) Regression graph G_{reg1}^N , b) a Markov equivalent regression graph G_{reg2}^N to G_{reg1}^N , c) a regression graph G_{reg3}^N that is directed acyclic and not Markov equivalent to G_{reg1}^N .

In G_{reg1}^N there are three collision Vs $(3, 4, 5)$, $(1, 2, 5)$, and $(2, 1, 3)$. In G_{reg2}^N there are the same collision Vs. Therefore, these two graphs are Markov equivalent. However, there are only two collision Vs in G_{reg3}^N these are $(3, 4, 5)$, and $(2, 1, 3)$. Hence this graph is not Markov equivalent to G_{reg1}^N and G_{reg2}^N . The Markov equivalence of the graphs in Figure 2 to the subgraph induced by $\{b, c\}$ in Figure 1 are further applications of Theorem 1. Notice that Propositions 3 to 8 of Section 6 result as special cases of Theorem 1.

The following algorithm generates a directed acyclic graph from a given G_{reg}^N that fulfills its known necessary conditions for Markov equivalence to a directed acyclic graph; see Proposition 2 of Wermuth (2010). We refer to these connected components as the blocks of G_{reg}^N .

Algorithm 1. (Obtaining a Markov equivalent directed acyclic graph from a regression graph). *Start from any given G_{reg}^N which has a chordal concentration graph and no chordless collision path in four nodes.*

1. *Apply the maximum cardinality search algorithm on the block consisting of full lines to order the nodes of the block.*
2. *Orient the edges of the block from a higher number to a lower one.*
3. *Replace collision Vs by sink Vs, i.e. replace $i\text{---}\circ\text{---}k$ and $i\text{---}\circ\leftarrow k$ by $i\rightarrow\circ\leftarrow k$ when i and k are uncoupled. When a dashed line in a block is replaced by an arrow, label the endpoints such that the arrow is from a higher number to a lower one if the labels do not already exist.*
4. *Replace dashed lines $i\text{---}\circ\text{---}k$ of triangles by a sink path $i\rightarrow\circ\leftarrow k$. When a dashed line in a block is replaced by an arrow, label the endpoints such that the arrow is from a higher number to a lower one if the labels do not already exist.*
5. *Replace dashed lines by arrows from a higher number to a lower one.*

Continually apply each step until it is not possible to continue applying it further. Then move to the next step.

Lemma 5. *For a regression graph with a chordal concentration graph and without chordless collision paths in four nodes, Algorithm 1 generates a directed acyclic graph that is Markov equivalent to G_{reg}^N .*

Proof. The generated graph is directed since by Algorithm 1, all edges are turned into arrows. Since the block containing full lines is chordal, the graph generated by the perfect elimination order of the maximal cardinality search does not have a directed cycle; see Blair and Peyton (1993) Section 2.4 and Tarjan and Yannakakis (1984).

In addition, the arrows present in the graph do not change by the algorithm. Thus, to generate a cycle containing an arrow of the original graph, there should have been a cycle in the directed graph generated by replacing blocks by nodes. But, this is impossible in a regression graph. Therefore in the generated graph, there is no cycle containing arrows that have been between the blocks of the original graph.

Within a block, all arrows point from nodes with higher numbers to nodes with lower ones. Otherwise, there would have been at step 3 of the algorithm a chordless collision path with four nodes in the graph. Hence no directed cycle can be generated.

Theorem 1 gives Markov equivalence to G_{reg}^N since Algorithm 1 preserves the skeleton of G_{reg}^N and no additional collision V is generated because sink oriented V s remain, only dashed lines are turned into arrows and no arrows are changed to dashed lines. \square

Notice that this algorithm does not generate a unique directed acyclic graph, but every generated directed acyclic graph is Markov equivalent to the given regression graph. To obtain the overall complexity of Algorithm 1, we denote by n the number of nodes in the graph and by e the number of edges in the graph.

Corollary 2. *The overall complexity of Algorithm 1 is $O(e^3)$.*

Proof. Suppose that the input of Algorithm 1 is a sequence of triples, each of which consists of the two endpoints of an edge and of the type of edge. The length of this sequence is equal to e and the highest number appearing in the sequence is n . For example, the sequence to the graph of Figure 15a) is $((1, 2, d), (3, 1, a), (5, 2, a), (4, 3, d), (4, 5, d))$, where ‘d’ corresponds to a dashed line and ‘a’ corresponds to an arrow pointing from the first entry to the second one. Notice that this labeling is in general not the same as the ordering of nodes given by Algorithm 1.

The first two steps of Algorithm 1 can be performed in $O(e + n)$ time; see Blair and Peyton (1993). Step 3 of Algorithm 1 may be performed in $e(e + 1)(e - 2)/2$ steps since for each edge, one can go through the edge set to find the edges that give a three node

path with an inner collision node. This needs $e(e + 1)/2$ steps. For each collision node, one goes again through the edge set, excluding the two edges involved in the collision path, to check if the collision is a \mathbf{V} . Other actions can be done in constant time.

Step 4 may require $ne(e + 1)/2$ steps since paths considered $\circ\text{---}\circ\text{---}\circ$ which do not form a \mathbf{V} . Therefore, there is no reason to go through the edge set for the third time, but one might need to go through the node ordering to decide on the direction of the generated arrow. The last step may be performed with ne steps by going through the edge set changing 'd's to 'a's appropriately by looking at the node ordering. Therefore, the overall complexity of Algorithm 1 is $O(e^3)$. \square

Corollary 2 and Propositions 4 to 8 can now be derived as special cases of Theorem 1 and Lemma 4. In addition by using Lemma 1, Lemma 2 and pairwise independences, subclasses of regression graphs can be identified which intersect with directed acyclic graphs, with other types of chain graphs, with concentration graphs or with covariance graphs.

Theorem 2. *A regression graph with a chordal graph for the context variables can be oriented to be Markov equivalent to a directed acyclic graph in the same skeleton, if and only if it does not contain any chordless collision path in four nodes.*

Proof. Every chordal concentration graph can be oriented to be equivalent to a directed acyclic graph; see Tarjan and Yannakakis (1984). A missing edge for node pair $i < k$ in a directed acyclic graph means $i \perp\!\!\!\perp k \mid > i \setminus k$ which would contradict 2(iii) if the graph contained a semi-directed chordless collision path in four nodes. No undirected chordless collision path in four nodes can be fully oriented without changing a collision \mathbf{V} into a transmitting \mathbf{V} , but G_{reg}^N can be oriented using Algorithm 1 if it contains no such path. \square

Notice that for joint Gaussian distributions, Theorem 2 excludes Zellner's seemingly unrelated regressions and it excludes covariance graphs that cannot be made Markov equivalent to fully directed acyclic graphs; see Proposition 4.

Proposition 8. *A multivariate regression graph with connected components g_1, \dots, g_J is an AMP chain graph in the same connected components if and only if the covariance graph of every connected component g_1, \dots, g_r of responses is complete.*

Proof. The conditional relations of the joint response nodes in an AMP chain graph coincide with those of the regression graph with the same connected components. Furthermore, the subgraph induced by each connected component g_j of an AMP chain graph is a concentration graph given $g_{>j}$ while in G_{reg}^N it is a covariance graph given $g_{>j}$. By Proposition 6, these have to be complete for Markov equivalence. \square

Proposition 9. *A multivariate regression graph with connected components g_1, \dots, g_J is a LWF chain graph in the same connected components if and only if it contains no semi-directed chordless collision path in four nodes and the covariance graph of every connected response component g_1, \dots, g_r is complete.*

Proof. The proof for the connected components of a LWF chain graph is the same as for an AMP chain graph since they both have concentration graphs for g_j given $g_{>j}$. The dependences of joint responses g_j on $g_{>j}$ coincide in a LWF chain graph with the bipartite part of the concentration graph in $g_j \cup g_{>j}$ so that Markov equivalent independence statements can only hold with these bipartite graphs being complete. \square

Figure 16 illustrates Propositions 2 to 9 with modified graphs of Figure 4.

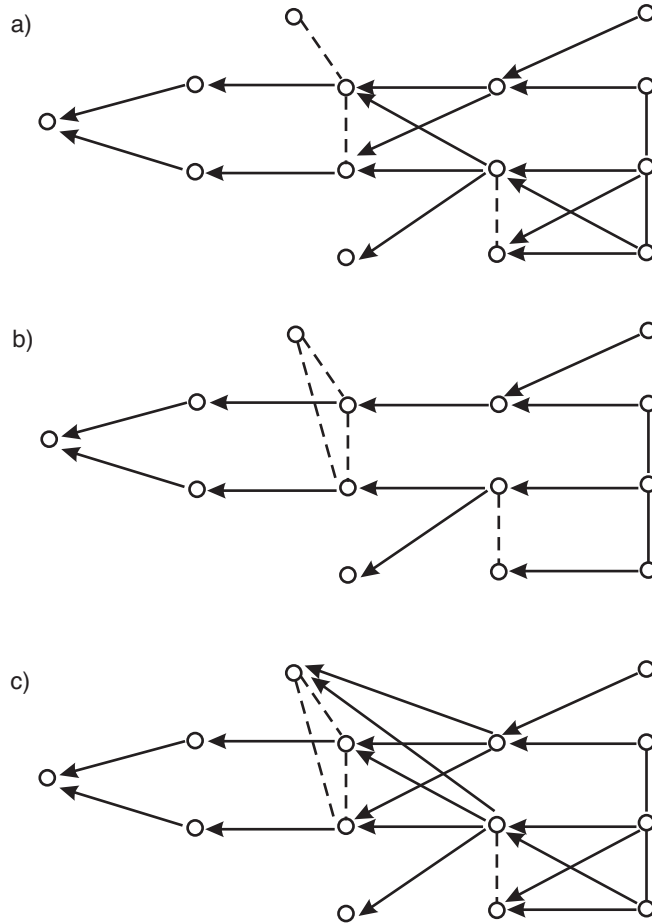


Figure 16: The graph of Figure 4 modified by adding edges to obtain a graph that is Markov equivalent to a) a directed acyclic graph b) an AMP chain graph in the same connected components c) a LWF chain graph in the same connected components.

The graphs in Figure 16 are Markov equivalent to a) a directed acyclic graph with the

same skeleton obtainable by Algorithm 1, b) an AMP chain graph in the same connected components and c) a LWF chain graph in the same connected components.

In general, by inserting some edges, a regression graph model can be turned into a model in one of the intersecting classes used in Propositions 2 to 9, just as a non-chordal graph may be turned into chordal one by adding edges. When the independence structure of interest is captured by an edge-minimal regression graph, then the resulting graph after adding edges will no longer be an edge-minimal graph and hence not give the most compact description possible.

However, the graph with some added edges may define a covering model that is easier to fit than the reduced model corresponding to the edge-minimal graph, just as an unconstrained Gaussian bivariate response regression on two regressors may be fitted in closed form, while the maximum-likelihood fitting in the reduced model of Zellner's seemingly unrelated regression requires iterative fitting algorithms. Any well-fitting covering model in the three intersecting classes will show weak dependences for the edges that are to be removed to obtain an edge-minimal graph.

Notice that sequences of regressions in the intersecting class with LWF chain graphs correspond for Gaussian distributions to sequences of the general linear models of Anderson (1958), that is to models in which each joint response has the same set of regressor variables. This shows in G_{reg}^N by identical sets of nodes from which arrows point to each node within a connected component.

In contrast, the models in the intersecting classes with the two types of undirected graph may be quite complex in the sense of including many merely generated chordless cycles of size four or larger.

Proposition 10. *A multivariate regression graph has the skeleton concentration graph if and only if it contains no collision \mathbf{V} and it has the skeleton covariance graph if and only if it contains no transmitting \mathbf{V} .*

Proof. Every \mathbf{V} is a collision \mathbf{V} in a covariance graph and a transmitting \mathbf{V} in a concentration graph implying contradictory independences; see Lemma 1, Lemma 2. \square

Lastly, Figure 17 shows the overall concentration graph induced by G_{reg}^N of Figure 4. It may be obtained from the given G_{reg}^N by finding first the smallest intersecting covering LWF chain graph in the same connected components, then closing every sink \mathbf{V} and finally changing all edges to full lines.

In such a graph, several chordless cycles in four or more nodes may be induced and the connected components of G_{reg}^N may no longer show. In such a case, much of the important structure of the generating regression graph is lost. In addition, merely induced chordless cycles require iterative algorithms for maximum-likelihood estimation,

even for Gaussian distributions. Thus, in the case of connected joint responses, it may be unwise to use a model search within the class of concentration graph models.

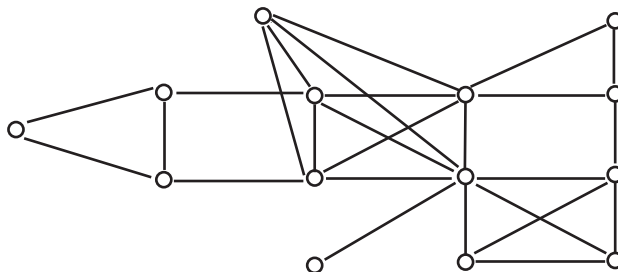


Figure 17: The overall concentration graph induced by the regression graph in Figure 4.

This contrasts with LWF chain graphs that coincide with regression graphs, such as in Figure 16c). These preserve the available prior knowledge about the connected components and give Markov equivalence to directed acyclic graphs so that model fitting is possible in terms of single response regressions, that is by using just univariate conditional densities. In addition, the simplified criteria for Markov equivalence of directed acyclic graphs apply.

On the other hand, sequences of regressions which coincide with LWF chains, permit also to model simultaneous intervention on a set of variables since the corresponding independence graphs are directed and acyclic in nodes representing vector variables. This represents a conceptually much needed extension of distributions generated over directed acyclic graphs in nodes representing single variables, but excludes the more specialized seemingly unrelated regressions and includes only complete covariance graphs.

Appendix: Details of regressions for the chronic pain data

The following tables show the results of linear least-squares regressions or logistic regressions, one at a time, for each of the response variables and for each component of a joint response separately. At first, each response is regressed on all its potentially explanatory variables given by their first ordering. The tables give the estimated constant term and for each variable in the regression, its estimated coefficient (coeff), the estimated standard deviation of the coefficient (s_{coeff}), as well as the ratio $z_{\text{obs}} = \text{coeff}/s_{\text{coeff}}$. These ratios are compared with 2.57, the 0.995 quantile of a random variable Z having a standard Gaussian distribution, for which $\Pr(Z > |2.57|) = 0.01$. In backward selection steps, the variable with the smallest observed value $|z_{\text{obs}}|$ is deleted from a regression equation, one at a time, until the threshold is reached.

This defines a selected model, unless one of the excluded variables has a contribution

of $|z'_{\text{obs}}| > 2.57$ when added alone to the selected directly explanatory variables, then such a variable needs to be included as an important directly explanatory variable. This did not happen in the given data set.

Response: Y , success of treatment; linear regression including a quadratic term							
explanatory variables	starting model			selected			excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
constant	23.40	–	–	20.50	–	–	–
Z_a , pain intensity after	–1.73	0.15	–11.19	–1.89	0.15	–12.77	–
X_a , depression after	–0.16	0.05	–3.04	–	–	–	–1.86
Z_b , pain intensity before	0.04	0.16	0.26	–	–	–	0.65
X_b , depression before	0.10	0.05	1.82	–	–	–	0.33
U , pain chronicity	–0.15	0.30	–0.51	–	–	–	–0.99
A , site of pain	–2.27	0.91	–2.48	–	–	–	–2.33
V , previous illnesses	0.19	0.11	1.76	–	–	–	1.24
B , level of schooling	–0.50	0.78	–0.64	–	–	–	–0.22

$(Z_a - \text{mean}(Z_a))^2$	0.18	0.23	3.41	0.23	0.05	4.28	–

$R^2_{\text{full}} = 0.54$ Selected model $Y : Z_a + Z_a^2$ $R^2_{\text{sel}} = 0.49$							

Response: Z_a , intensity of pain after treatment; linear regression							
explanatory variables	starting model			selected			excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
constant	2.74	–	–	2.98	–	–	–
Z_b , pain intensity before	0.12	0.08	1.60	0.16	0.07	2.16*	–
X_b , depression before	0.03	0.02	1.28	–	–	–	1.76
U , pain chronicity	0.11	0.14	0.75	–	–	–	1.43
A , site of pain	1.07	0.42	2.51	1.27	0.39	3.26	–
V , previous illnesses	0.00	0.05	0.03	–	–	–	0.83
B , level of schooling	–0.19	0.37	–0.52	–	–	–	–0.70

$R^2_{\text{full}} = 0.09$ Selected model $Z_a : Z_b + A$ $R^2_{\text{sel}} = 0.07$							

*: depression before treatment is selected because of repeated measurement, the the low correlation is due to a change in measuring, before and after treatment

Response: X_a , depression after treatment; linear regression							
explanatory variables	starting model			selected			excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
constant	2.54	–	–	4.55	–	–	–
Z_b , pain intensity before	–0.05	0.22	–0.23	–	–	–	–0.21
X_b , depression before	0.62	0.06	10.43	0.68	0.05	12.68	–
U , pain chronicity	0.96	0.42	2.28	–	–	–	2.31
A , site of pain	–1.19	1.25	–0.95	1.27	0.39	3.26	–0.10
V , previous illnesses	0.05	0.15	0.35	–	–	–	1.08
B , level of schooling	–0.15	1.09	0.14	–	–	–	–0.01
$R^2_{\text{full}} = 0.46$ Selected model $X_a : X_b$ $R^2_{\text{sel}} = 0.45$							

The tables show for linear models also R^2 , the coefficient of determination, both for the full and for the selected model. Multiplied by 100, it gives the percentage of the variation in the response explained by the model.

In the linear regression of Z_a on X_a and on Z_b, A, X_b , the directly explanatory variables of both Z_a and X_a , the contribution of X_a leads to $z_{\text{obs}} = 3.40$ which coincides – by definition – with z_{obs} computed for the contribution of Z_a in the linear regression of X_a on Z_a and on the directly explanatory variables of both Z_a and X_a . Hence the two responses are correlated even after considering the directly explanatory variables and a dashed line joining Z_a and Z_b is added to the well-fitting regression graph in Figure 8.

Response: Z_b , intensity of pain before; linear regression							
explanatory variables	starting model			selected			excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
constant	7.60	–	–	7.38	–	–	–
U , pain chronicity	0.10	0.13	0.77	–	–	–	0.59
A , site of pain	–0.58	0.40	–1.44	–	–	–	–1.20
V , previous illnesses	0.02	0.05	0.46	–	–	–	0.72
B , level of schooling	–0.94	0.35	–2.70	–0.89	0.22	–2.65	–
$R^2_{\text{full}} = 0.05$ Selected model $Z_a : B$ $R^2_{\text{sel}} = 0.03$							

Response: X_b , depression before; linear regression							
explanatory variables	starting model			selected			excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
constant	10.96	—	—	7.31	—	—	—
U , pain chronicity	1.97	0.49	4.02	1.78	0.46	3.87	—
A , site of pain	-2.33	1.50	-1.55	—	—	—	-1.42
V , previous illnesses	0.54	0.18	2.99	0.55	0.18	3.06	—
B , level of schooling	-1.10	1.31	-0.84	—	—	—	-0.57
$R^2_{\text{full}} = 0.18$ Selected model $X_b : U + V$ $R^2_{\text{sel}} = 0.17$							

In the linear regression of Z_b on X_b and on U, A, B , the directly explanatory variables of both Z_b and X_b , the contribution of X_b leads to $z_{\text{obs}} = 2.64$ which coincides with z_{obs} computed for the contribution of Z_b in the linear regression of X_b on Z_a, U, A, B . Hence the two responses are associated after considering their directly explanatory variables and there is a dashed line joining Z_b and X_b in the regression graph of Figure 8.

Response: U , chronicity of pain; linear regression							
explanatory variables	starting model			selected			excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
constant	2.93	—	—	2.47	—	—	—
A , site of pain	0.95	0.21	4.58	1.02	0.20	5.02	—
V , previous illnesses	0.14	0.02	5.83	0.14	0.02	5.92	—
B , level of schooling	-0.27	0.19	-1.43	—	—	—	-1.43
$R^2_{\text{full}} = 0.26$ Selected model $X_b : A + V$ $R^2_{\text{sel}} = 0.25$							

Response: A , site of pain; logistic regression							
explanatory variables	starting model			selected			excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
constant	0.26	—	—	0.52	—	—	—
V , previous illnesses	0.05	0.04	1.22	—	—	—	1.22
B , level of schooling	-1.25	0.40	-3.11	-1.28	0.40	-3.18	—
Selected model $A : B$							

Response: V , previous illnesses; linear regression							
explanatory variables	starting model			selected			excluded
	coeff	s_{coeff}	z_{obs}	coeff	s_{coeff}	z_{obs}	z'_{obs}
constant	6.41	—	—	5.33	—	—	—
B , level of schooling	-0.65	0.54	-1.20	—	—	—	—

Selected model V : —

The relatively strict criterion, for excluding variables, assures that all edges in the derived regression graph correspond to dependences and associations that are considered to be substantive in the given context. Had instead a 0.975 quantile been chosen as threshold, then one arrow from A to Y and another from U to X_a would have been added to the regression graph. Though this would correspond to a better goodness-of-fit, such weak dependences are less likely to become confirmed as being important in follow-up studies.

The subgraph induced by Z_a, Z_b, X_a, X_b of the regression graph in Figure ?? corresponds to two seemingly unrelated regressions, a term coined by Zellner (1962). In this situation, separate least-squares regressions can lead to estimates that differ for joint Gaussian distributions much from the proper maximum-likelihood estimates (m.l.e) because the responses are correlated; see (Havelmoo, 1943). There may even exist several solutions to the likelihood equations; see Drton and Richardson (2004).

However for the present data, the structure is so evidently well-fitting in a covering model that none of these potential problems are relevant, that is in the unconstrained multivariate regression of Z_a and X_a on Z_b, X_b, U, V, A, B .

With $C = \{U, V, A, B\}$, this is evident from the observed covariance matrix of Z_a, X_a given Z_b, X_b, C , denoted here by $\tilde{\Sigma}_{aa|bC}$ and the observed regression coefficient matrix $\tilde{\Pi}_{a|b.C}$ being almost identical to the corresponding m.l.e $\hat{\Sigma}_{aa|bC}$ and $\hat{\Pi}_{a|b.C}$.

The former can be obtained by sweeping or partially inverting the observed covariance matrix of the eight variables with respect to Z_b, X_b, C and the latter by using an adaption of the EM-algorithm, due to Kiiveri (1989), on the observed covariance matrix of the four symptoms, corrected for linear regression on C . In this way, one gets

$$\tilde{\Sigma}_{aa|bC} = \begin{pmatrix} 5.61 & 3.91 \\ 3.91 & 48.37 \end{pmatrix}, \quad \hat{\Sigma}_{aa|bC} = \begin{pmatrix} 5.66 & 3.94 \\ 3.94 & 48.41 \end{pmatrix},$$

$$\tilde{\Pi}_{a|b.C} = \begin{pmatrix} 0.12 & 0.03 \\ -0.05 & 0.62 \end{pmatrix}, \quad \hat{\Pi}_{a|b.C} = \begin{pmatrix} 0.14 & 0.00 \\ 0.00 & 0.60 \end{pmatrix}.$$

The assumed definition of the joint distribution in terms of univariate and multivariate regressions assures that the overall fit of the model can be judged locally in two steps. First, one compares each unconstrained, full regression of a single response with regressions constrained by some independences, that is by selecting a subset of directly explanatory variables from the list of the potentially explanatory variables. Next, one decides for each component pair of a joint response whether this pair is conditionally independent given their directly explanatory variables considered jointly. This can again be achieved by single univariate regressions, as illustrated above for Z_a and X_a .

Acknowledgement. The work of the second author has been supported in part by the Swedish Research Society via the Gothenburg Stochastic Center and by the Swedish Strategic Fund via the Gothenburg Mathematical Modelling Center. We thank D.R. Cox and R. Castelo for their helpful comments and discussions.

References

- Ali, R.A., Richardson, T.S. and Spirtes, P. (2009). Markov equivalence for ancestral graphs. *Ann. Statist.* **37**, 2808–2837.
- Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. (3rd ed., 2003) Wiley, New York.
- Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist.* **1**, 135–141.
- Andersson, S.A. and Perlman, M.D. (2006). Characterizing Markov equivalence classes for AMP chain graph. *Ann. Statist.* **34**, 939–972.
- Andersson, S.A., Madigan, D., and Perlman, M.D. (2001). Alternative Markov properties for chain graphs. *Scand. J. Statist.* **28**, 33–86.
- Andersson, S.A., Madigan D., Perlman M.D. and Triggs, C.M. (1997). A graphical characterization of lattice conditional independence models. *Ann. Math. Artif. Intell.* **21**, 27–50.
- Barndorff-Nielsen, O.E. (1978). *Information and exponential families in statistical theory*. Wiley, Chichester.
- Bergsma, W. and Rudas, T. (2002). Marginal models for categorical data. *Ann. Statist.* **30**, 140–159.
- Birch, M.W. (1963). Maximum likelihood in three-way contingency tables. *J. Roy. Statist. Soc. B* **25**, 220–233.
- Bishop, Y.M.M., Fienberg, S.F. and Holland, P.W. (1975). *Discrete multivariate analysis*. MIT Press, Cambridge.

- Blair, J. R. S. and Peyton, B. W. (1993). An introduction to chordal graphs and clique trees. In: *Graph theory and sparse matrix computations*, IMA Volumes in mathematics and its applications, Vol. 56. (eds. J. A. George, J. R. Gilbert, and J. W. H. Liu) Springer, New York, 1–30.
- Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley, New York.
- Brown, L.D. (1986). *Fundamentals of statistical exponential families with applications in statistical decision theory*. LNMS **9**, Inst. Math. Statist., Beachwood.
- Castelo, R. and Kocka, T. (2003). On inclusion-driven learning of Bayesian networks. *J. Machine Learn. Res.* **4**, 527–574.
- Castelo, R. and Siebes, A. (2003). A characterization of moral transitive acyclic directed graph Markov models as labeled trees. *J. Stat. Plan. Inf.* **115**, 235–259.
- Caussinus, H. (1966). Contribution à l’analyse statistique des tableaux de corrélation. *Annales de la Faculté des Sciences de l’Université de Toulouse* **29**, 77–183.
- Cayley, A. (1889). A theorem on trees. *Quart. J. Math.* **23**, 376–378.
- Chaudhuri, S., Drton, M. and Richardson, T.S. (2007). Estimation of a covariance matrix with zeros *Biometrika* **94**, 199–216.
- Cochran, W. G. (1938). The omission or addition of an independent variate in multiple linear regression. *Suppl. J. Roy. Statist. Soc.* **5**, 171–176.
- Chickering, D.M. (1995). A transformational characterization of equivalent Bayesian networks. In *Proc. 10th UAI conf.* (eds. P. Besnard and S. Hanks) Morgan Kaufman, San Mateo, 87–98.
- Cox, D. R. (1966). Some procedures associated with the logistic qualitative response curve. In *Research papers in statistics: Festschrift for J. Neyman* (ed. F.N. David), Wiley, New York, 55–71.
- Cox, D.R. (2006). *Principles of statistical inference*. Cambridge University Press, Cambridge.
- Cox, D.R. and Wermuth, N. (1990). An approximation to maximum-likelihood estimates in reduced models. *Biometrika* **77**, 747–761.
- Cox, D.R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statist. Science* **8**, 204–218; 247–277.
- Cox, D.R. and Wermuth, N. (1996). *Multivariate dependencies: models, analysis, and interpretation*. Chapman and Hall (CRC), London.
- Cox, D.R. and Wermuth, N. (2003). A general condition for avoiding effect reversal after marginalization. *J. Roy. Statist. Soc. B* **65**, 937–941.
- Darroch, J.N. (1962). Interactions in multi-factor contingency tables. *J. Roy. Statist. Soc. B* **24**, 251–263.
- Darroch, J.N., Lauritzen, S.L. and Speed, T.P. (1980). Markov fields and log-linear models for contingency tables. *Ann. Statist.* **8**, 522–539.
- Dawid, A.P. (1979). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. B* **41**, 1–31.

- Dempster, A.P. (1969). *Elements of continuous multivariate analysis*. Addison-Wesley, Reading, Mass.
- Dempster, A.P. (1972). Covariance selection. *Biometrics* **28**, 157–175.
- Dinitz, Y. (2006). Dinitz’ algorithm: the original version and Even’s version. In: *Essays in memory of Shimon Even*. (eds. S. Even, O. Goldreich, A.L. Rosenberg, and A.L. Selman) Springer, New York, 218–240.
- Dirac, G.A. (1961). On rigid circuit graphs. *Abhandl. Math. Seminar Hamburg* **25**, 71–76.
- Drton, M. (2009). Discrete chain graph models. *Bernoulli* **15**, 736–753.
- Drton, M., Eichler, M. and Richardson, T.S. (2009). Computing maximum likelihood estimates in recursive linear models. *J. Mach. Learn. Res.* **10**, 2329–2348.
- Drton, M. and Perlman, M.D. (2004). Model selection for Gaussian concentration graphs. *Biometrika* **91**, 591–602.
- Drton, M. and Richardson, T.S. (2004). Multimodality of the likelihood in the bivariate seemingly unrelated regression model. *Biometrika* **91**, 383–392.
- Drton, M. and Richardson, T.S. (2008a). Binary models for marginal independence. *J. Roy. Statist. Soc. B* **70**, 287–309.
- Drton, M. and Richardson, T.S. (2008b). Graphical methods for efficient likelihood inference in Gaussian covariance models. *J. Machine Learn. Res.* **9**, 893–914.
- Edwards, D. (2000). *Introduction to graphical modelling*. (2nd ed.) Springer, New York.
- Frydenberg, M. (1990). The chain graph Markov property. *Scand. J. Statist.* **17**, 333–353.
- Geiger, D., Verma, T.S. and Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks* **20**, 507–534.
- Glonek G.F.V. and McCullagh P. (1995). Multivariate logistic models. *J. Roy. Statist. Soc. B* **53**, 533–546.
- Goodman, L.A. (1970). The multivariate analysis of qualitative data: interaction among multiple classifications. *J. Amer. Statist. Assoc.* **65**, 226–256.
- Hardt, J., Sidor, A., Nickel, R., Kappis, B., Petrak, F., and Egle, U. T. (2008). Childhood adversities and suicide attempts: a retrospective study. *J. Family Violence*, **23**, 713–718.
- Holland P.W. and Rosenbaum, P.R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Ann. Statist.* **14**, 1523–1543.
- Jensen, S.T. (1988) Covariance hypotheses which are linear in both the covariance and the inverse covariance. *Ann. Statist.* **16**, 302–322.
- Jöreskog, K.G (1981). Analysis of covariance structures. *Scand J Statist.* **8**, 65–92.
- Kang, C. and Tian, J. (2009). Markov properties for linear causal models with correlated errors. *J. Mach. Learn. Res.* **10**, 41–70.
- Kappesser, J. (1997). *Bedeutung der Lokalisation für die Entwicklung und Behandlung chronischer Schmerzen*. Thesis, Department of Psychology, University of Mainz.

- Kauermann, G. (1996). On a dualization of graphical Gaussian models. *Scand. J. Statist.* **23**, 115–116.
- Kiiveri, H.T. (1987). An incomplete data approach to the analysis of covariance structures. *Psychometrika* **52**, 539–554.
- Kiiveri, H.T., Speed, T.P. and Carlin, J.B. (1984). Recursive causal models. *J. Austral. Math. Soc A* **36**, 30–52.
- Kline, R.B. (2006). *Principles and practice of structural equation modeling* (3rd edition). Guilford Press, New York.
- Koster, J. (2002). Marginalising and conditioning in graphical models. *Bernoulli* **8**, 817–840.
- Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press, Oxford.
- Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for association between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17**, 31–57.
- Lehmann, E.L. and Scheffé, H. (1955). Completeness, similar regions and unbiased estimation. *Sankhya* **14**, 219–236.
- Levitz, M., Perlman, M. D., & Madigan, D. (2001). Separation and completeness properties for AMP chain graph Markov models. *Ann. Statist.* **29**, 1751–1784.
- Lněnička, R. and Matúš, F. (2007). On Gaussian conditional independence structures. *Kybernetika* **43**, 323–342.
- Lupparelli M., Marchetti, G. M., and Bergsma, W. P. (2009). Parameterization and fitting of discrete bi-directed graph models. *Scand. J. Statist.* **36**, 559–576.
- Ma, Z.M., Xie, X.C. and Geng, Z. (2006). Collapsibility of distribution dependence. *J. Roy. Statist. Soc. B* **68**, 127–133.
- Mandelbaum, A. and Rüschendorf, L. (1987). Complete and symmetrically complete families of distributions. *Ann. Statist.* **15**, 1229–1244.
- Marchetti, G.M. and Lupparelli, M. (2010). Chain graph models of multivariate regression type for categorical data. *Bernoulli*, to appear.
- Marchetti, G.M. and Wermuth, N. (2009). Matrix representations and independencies in directed acyclic graphs. *Ann. Statist.* **47**, 961–978.
- McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall (CRC), London.
- Nelder, J.A. and Wedderburn, R. (1972). Generalized Linear Models. *J. Roy. Statist. Soc. A* **135**, 37–384.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann, San Mateo.
- Pearl, J. (2009). *Causality: models, reasoning, and inference*. 2nd ed., Cambridge University Press, New York.
- Pearl J. and Paz, A. (1987). Graphoids: a graph based logic for reasoning about relevancy revelations. In: (eds. B.D. Boulay D. Hogg, and L. Steel) *Advances in artificial intelligence II*, North Holland, Amsterdam, 357–363.

- Pearl, J. and Wermuth, N. (1994). When can association graphs admit a causal interpretation? In: *Models and data, artificial intelligence and statistics IV.* (eds. P. Cheeseman and W. Oldford). Springer, New York, 205–214.
- Richardson, T.S. and Spirtes, P. (2002). Ancestral Markov graphical models. *Ann. Statist.* **30**, 962–1030.
- Roverato, A. (2005). A unified approach to the characterisation of Markov equivalence classes of directed acyclic graphs, chain graphs with no flags and chain graphs. *Scand. J. Statist.* **32**, 295–312.
- Roverato, A. and Studený M. (2006). A graphical representation of equivalence classes of AMP chain graphs. *J. Mach. Learn. Res.* **7**, 1045–1078.
- Rudas, T., Bergsma W. P. and Nemeth, R. (2010). Marginal log-linear parameterization of conditional independence models. *Biometrika*, to appear.
- Sadeghi, K. (2009). Representing modified independence structures. *Transfer thesis, Oxford University.*
- San Martin E., Mochart M. and Rolin, J.M. (2005). Ignorable common information, null sets and Basu’s first theorem. *Sankhya* **67**, 674–698.
- Speed, T.P. and Kiiveri, H.T. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* **14**, 138–150.
- Spirtes, P., Glymour C. and Scheines R. (1993). *Causation, prediction and search.* Springer, New York.
- Stanghellini, E. and Wermuth, N. (2005). On the identification of path analysis models with one hidden variable. *Biometrika* **92**, 337–350.
- Studený, M. (2005). *Probabilistic conditional independence structures.* Springer, London.
- Sundberg, R. (2010). Flat and multimodal likelihoods and model lack of fit in curved exponential families. *Scand. J. Statist.*, to appear.
- Tarjan, R.E. and Yannakakis, M. (1984). Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs *SIAM J. Comp* **13**, 566 – 579.
- Verma, T. and Pearl J. (1990). Equivalence and synthesis of causal models. Proc. 6th UAI conf. (eds. P.P. Bonissone, M. Henrion, L.N. Kanal and J.F. Lemmer). Elsevier, Amsterdam, 220–227.
- Wermuth, N. (1976a). Analogies between multiplicative models for contingency tables and covariance selection. *Biometrics* **32**, 95–108.
- Wermuth, N. (1976b). Model search among multiplicative models. *Biometrics* **32**, 253–263.
- Wermuth, N. (1980). Linear recursive equations, covariance selection, and path analysis. *J. Amer. Statist. Assoc.* **75**, 963–97.
- Wermuth, N. (2011). Probability models with summary graph structure. *Bernoulli*, to appear.

- Wermuth, N. and Cox, D.R. (1998). On association models defined over independence graphs. *Bernoulli* **4**, 477–495.
- Wermuth, N. and Cox, D.R. (2004). Joint response graphs and separation induced by triangular systems. *J.Roy. Stat. Soc. B* **66**, 687-717.
- Wermuth, N. and Cox, D.R. (2008). Distortions of effects caused by indirect confounding. *Biometrika* **95**, 17–33.
- Wermuth, N., Cox, D.R. and Marchetti, G.M. (2006). Covariance chains. *Bernoulli* **12**, 841-862.
- Wermuth, N. and Lauritzen, S.L. (1983). Graphical and recursive models for contingency tables. *Biometrika* **70**, 537–552.
- Wermuth, N. and Lauritzen, S.L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *J. Roy. Statist. Soc. B* **52**, 21–75.
- Wermuth, N., Marchetti, G.M. and Cox, D.R. (2009). Triangular systems for symmetric binary variables. *Electr. J. Statist.* **3**, 932–955.
- Wermuth, N., Wiedenbeck, M. and Cox, D.R. (2006). Partial inversion for linear systems and partial closure of independence graphs. *BIT, Numerical Mathematics* **46**, 883–901.
- Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Wiley, Chichester.
- Wiedenbeck, M. and Wermuth, N. (2010). Changing parameters by partial mappings. *Statistica Sinica* **20**, 823–836.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* **57**, 348–368.
- Zhao, H., Zheng, Z. and Liu, B. (2005). On the Markov equivalence of maximal ancestral graphs. *Science China A* **48**, 548–562.